Tenzing Briggs

6 October 2023

Comparative Review of Top 10 Google Scholar Results for "Identifying Influential Bloggers": What Do Differences in Studies Show Us About Researching the Blogosphere?

Introduction

As a comparative review, this paper seeks to find the similarities and differences which consistently appear across the given top 10 sources on the topic "Identifying Influential Bloggers" when it comes to their individual goals, approaches, and findings. Yet, on a deeper level, this paper also attempts to then compare said similarities/differences *across* the goals, approaches, and findings; that is, can comparative review show how common goals lead to common approaches, and then to common findings? Put another way, this comparative review looks at the common issues in goals that then lead to common issues between approaches, as well as common issues between findings that stem either from what is shared or different in approach.

Goals

All sources came from searching for "Identifying Influential Bloggers", so obviously the most basic shared goal is that all seek to identify influential bloggers, albeit via different models/metrics. Similarly, all sources thus mention and theorize about the nature of the blogosphere. Yet, there are more implicit shared goals I feel worth mention, as they tell us the importance of influential bloggers. Namely, I see some common trends across the Abstracts, Introductions, and Related Works sections of these 10 sources as a starting point for implicit goals in description of the blogosphere. A majority of these sources mention in these places the "real-world" where we also "consult others" as a corollary to how the blogosphere also can have

real-world impact through providing sources we consult online for decision making (Agarwal et al., 2008; Akritidis e al., 2009; Khane et al., 2015; Aziz & Rafi, 2010; Ishfaq et al., 2017). Following this, this implied "real-world" behind the blogosphere implies also the goal of relating it to explicit real-world applications; this is implied across all Introductions and Abstracts that mention how influential actors can be utilized.

For example, all sources mention the commercial applications (Agarwal et al., 2008; Akritidis et al., 2009; Akritidis et al., 2011; Aziz & Rafi, 2010; Ishfaq et al., 2017; Kayes et al., 2012; Khan et al., 2015; Khan et al., 2017; Moh & Shola, 2013), where influencers are 'market movers.' All also mention the ability of influential bloggers to 'sway,' opinion, but, meaningfully, a majority (although not all) also relate influentials more explicitly to politics (Khan et al., 2017; Akritidis et al., 2009; Akritidis et al., 2011; Khan et al., 2015; Kayes et al., 2012; Moh & Shola, 2013; Agarwal et al., 2012; Aziz & Rafi, 2010) or to citizen journalism (Kayes et al., 2012; Khan et al., 2017; Akritidis et al., 2009). And only two explicitly mention influential bloggers as showing trends in education and bibliometrics (Agarwal et al., 2012; Ishfaq et al., 2017). This shows there are accepted common applications, like marketing and social or political influence, even as there are more uncommon applications like bibliometrics and education; this could be considered as showing that influential bloggers can fulfill many goals, those both obvious and not.

There is also a shared goal of defining and addressing the challenges to quantitative analysis. This is evident in that almost all models look at how we can turn metrics into specific concepts (e.g. inlinks into a sign of authority), such in places where authors define the "intuitive" features of influential blogging (Agarwal et al., 2008). Note here that this shared goal is important because these metric-concepts can still be disputed, as we see in later approaches. E.g.,

Agarwal et al.'s (2008) conception of higher outlinks as a negative sign of novelty is disputed by Akritidis et al.'s (2009) conception of outlinks as "more subtle" (p. 78), where they "argue that the outlinks are not relevant to the post's novelty, and all links should have a single semantic, that of implying endorsement" (Akritidis et al., 2009, p. 78).

Additionally, all sources have the goal of furthering quantitative models with respect to others' related research. This is seen in the high self-referentiality among sources; almost all 10 sources all cited each other, in order of appearance: i.e, Akritidis et al., (2009) cites Agarwal et al. (2008); Akritidis et al. (2011) cites Akritidis et al. (2009); Khan et al., (2015) cites both Agarwal et al. (2008) and Akritidis et al. (2009); and so on. This is also seen in the common practice of comparing one's methods and models to similar ones by other authors; many models take an established model, and add a novel parameter or metric to try to improve rankings in a certain way, which we see in the Akritidis et al. (2009) time parameter, the Moh & Shola (2010) FBCount parameter, the Ishfaq et al. (2017) unique commenter and sentiment feature parameters, or the Khan et al. (2015) activity, activeness, consistency, and NormalizedPostLength parameters.

Approaches

First, a common approach is to use influence ranking/index models as a basis of judging influentials, attempting to identify the top *k* bloggers in a given blogger network. Researchers' approaches, then, start by defining the network, finding a way to rank its users according to the data of the hypothesized network, and evaluating that against an actual, real-world blog.

Second, all approaches start with some sort of classification, which is a general way of saying that they start by defining certain concepts to arrive at later approaches, like experimental design. Even in Khan et al. (2017), which lacks experimental design in the traditional sense as it

is rather a survey, classification is important for approaching the issue, such as their distinctions of models as feature-based versus network-base and as non-temporal versus temporal. Feature-based models focus on metadata features, like inlinks and number of comments, and classify what these features translate to, such as inlinks as recognition or comments as activity generation as in Agarwal et al. (2008). Likewise, network-based models, rather than focusing on individual blog posts' metadata, look at constructing a network graph of the blogger connections to further arrive at quantitative approaches, like deciding to focus on centrality measures. Notably, Kayes et al. (2012) is the only source to use this kind of model, with their six node centrality measures, but Khan et al. (2017) also informs us that there numerous other kinds of network-based models like CR algorithm, MIV, Interest vector, TDIR, and many others.

Third, parameters are a common theme which differ across sources even as certain ones are relatively constant, or at least always mentioned in their Related Research sections. For feature-based models, all sources also begin with the baseline parameters outlined in Agarwal et al. (2018): inlinks (recognition), number of comments (activity generation), outlinks (novelty), and post length (eloquence). Note, however, that the mention of these parameters sometimes leads to approaches which differ in that they *contrast* the original defined parameters in Agarwal et al. (2018) by seeking other ways of defining these parameters. Most notably, Akritidis et al. (2009) contests the idea that higher outlinks are a sign of negative novelty and puts forward the idea that high outlinks could be interpreted as endorsement.

Alongside this, many such models then all look at more complex or novel parameters that could be used with the original four. For example, some other more complex parameters seen are the following: productivity (Akritidis et al., 2009) and its cousins activity, activeness, and consistency (Khan et al., 2015); sentiment via TFIDF machine learning (Moh & Shola, 2013);

rate of comments (Agarwal et al., 2012); Semantic Similarity Measure, Sentence-Wise Length, and Sentence-Based Similarity Measure (Aziz & Rafi, 2010); and unique commenters (Ishfaq et al., 2017). New studies also commonly combine these new parameters in new ways, too, like making modules that combine parameters in a novel way, such as in Khan et al. (2015) with their productivity score, popularity score, and blog quality score modules.

Another common theme, but that is contested among specific sources, is how and why to use weights with parameters. Agarwal et al. (2018) justifies parameter weights as a way to either tune their model, seek stabilization of rankings with varied weights, or use weights to change the influence flow model in such a way so as to look at how its rankings change with weights of 0 (i.e. investigating the changes occuring when parameters are effectively removed). Interestingly, Akritidis et al. (2009) attempts to do away with weighted parameters under the critique that weights should be avoided so that other researchers reusing a given model do not need to retune or stabilize rankings. Yet, in Akritidis et al. (2011) weights reappear in their continued models, but for the different purpose of (rather than being used for tuning) using constant weights to prioritize certain parameters over others, such as inlinks and comments over outlinks.

Fourth, experimental design, like parameters, are of course part of every source, and what they consistently share are using data collection of specific datasets to try and confirm the hypotheses behind their given or proposed influence models. All use the approach of using experiments with real-world blog site datasets to confirm predictions, although in Khan et al. (2017) their survey's "dataset" is the corpus of research in the field. Because of the need for efficient and quick data collection, many datasets reappear across the sources due to their ease of collection, especially TUAW (Agarwal et al., 2008; Agarwal et al., 2012; Akritidis et al., 2009;

Khan et al., 2015) and Engadget (Akritidis et al., 2011; Moh & Shola, 2013; Agarwal et al., 2012). Khan et al. (2017) also identifies datasets commonly used.

With this comes the consistent use of the data to construct rankings as either per baseline, per their given proposed model, or occasionally per other models that they are seeking to emulate or critique. In short, common datasets mean researchers can easily compare their work to others. It means models can compare results of different ranking models, like comparing between rankings what different bloggers are 1st, 2nd, or 3rd ranked, or by using quantitative evaluation measures like OSim, Spearman's correlation, or Kendall's correlation. However, some datasets are used only once among these 10 sources, like BlogCatalog (Kayes et al., 2012), which at the time of their writing had symmetrical relationships and thus allowed their research to use an undirected graph; this, as well as the datasets described by Khan et al. (2017), could imply that different datasets have different advantages to research.

Findings

The findings in Khan et al. (2017) help orient us to common themes in the other 9 sources' findings, since their study is a survey of influential blogger studies in general. Khan et al. (2017) begins by showing that "a number of recent research work has provided some indirect ways to measure the correctness and accuracy of the model" (p. 81). At the same time, influence ranking models often share more direct ways of arriving at findings, such as evaluating different model rankings with techniques like OSim, Spearman's correlation, and Kendall correlation coefficient, which are all used by Akritidis et al. (2011), Khan et al. (2015), and Ishfaq et al. (2017). Khan et al. (2017) also identifies common deficiencies among findings, like the current lack of comparative analysis of blogs with other social media like YouTube, Flicker, LinkedIn (at least for their sources surveyed and for these 10 compared in this review). Similarly, Khan et al.

(2017) identifies "a need for topic-specific identification of influential bloggers" (p. 81). Most meaningfully, Khan et al. (2017) finds that comparing feature-based versus network-based models identifies salient feature- and network-based influences.

A common trend in findings amongst the other nine sources is the comparison of active versus influential bloggers, even as findings vary. The MEIBI and MEIBIX models of Akritidis et al. (2009) find that activity matters; these models account for recency or age of posts and their inlinks, and they claim these models improved the rankings by thus identifying the *now-influential* bloggers that are both active and influential. In their continued research, Akritidis et al. (2011) finds that their BP and BI indexes taken together can help characterize bloggers as *either* or *both* influential or recently productive (or neither). Similarly, the findings of Khan et al. (2015) match these former two studies, both because they found their "proposed methods identify the influential bloggers in a more effective manner" (p. 14) and because those methods account for productivity. In contrast, Agarwal et al. (2008), for example, makes a point of finding that there are influential bloggers that *aren't* active. Clearly, activeness matters, and most studies comment upon it in their findings, regardless of how activeness is valued by the given researchers.

Alongside this is the common findings that demonstrate the importance of temporal aspect—an aspect connected to whether a blogger is active—like the Akritidis et al. (2009) MEIBI and MEIBIX model findings that confirm that recent posts or recent inlinks and comments predict influence. Akritidis et al. (2009) also emphasizes how this inclusion can show "significant temporal patterns" (p. 83), and even Agarwal et al. (2008) and Agarwal et al. (2012), which both choose to find influentials that are influential across time (i.e. inactive influentials),

comment on temporal patterns when they describe the discovered categories of long-term, average-term, transient, and burgeoning influentials.

Because ground truth is absent in this topic due to blogosphere complexity, it also follows that a common approach is to try to use alternative methods other than ground truth to confirm hypothesized influence models. Thus, whenever a model is confirmed accurate, there is a common discovery of models and their criteria that can be considered going forward as reliable methods and as new baselines in expanded research.

However, findings can vary among authors based on the aspects their models focus upon. It has to be recognized that several models have vastly different findings. If they are all viable research, this points to the fact that different findings relate to their different approaches, which is important because it shows that approaches with novel metrics will likely also have novel findings for blogging influence aspects previously not considered. For example, Khan et al. (2015) found their MIIB model showed the importance of considering blog site importance; that is, their model was the only of the 10 to consider blog site importance, and so their findings (regardless of other author models and findings) especially inform blog site importance.

At the same time, despite different findings, authors consistently arrive at findings through similar evaluations of baseline and/or compared models, such as the previously mentioned common approaches of OSim, Spearman's correlation, and Kendall's correlation seen in Akritidis et al. (2011), Khan et al. (2015), and Ishfaq et al. (2017). In line with Khan et al. (2017), this could show how different models are all somewhat valid in their own way, since each different model accounts for and assigns different weights to different metrics, some of which are also novel metrics. Kayes et al. (2012), for example, uses network centrality measures and thus can comment best on finding the nature of the whole network, seen in their finding that

BlogCatalog classifies as a core-periphery structure. Moh & Shola (2013), by including many metrics and FBCount, create a model with higher differentiation due to its added focus on Facebook likes and shares.

Finally there should also be some recognition that findings differ because of the different approaches regarding parameters, especially parameter weights. The different findings across studies tells us that careful consideration of parameter emphasis via weights is necessary, as our findings will fluctuate depending on what parameters we value most in our model via use of weights.

Conclusion

In conclusion, these 10 sources demonstrate the line between common and different practices. Commonality mixes with difference, in that researchers can simultaneously approach the common issues in identifying influential bloggers both from similar practices while from different standpoints. Underneath the overarching goal of influential blogger identification, certain other goals lie implicit: bloggers should be identified due to their influence on the real-world, and to achieve this the blogosphere and its members must be defined, classified, and studied in conjunction with other researchers and their work.

This basis of the issue means researchers start their approach by defining the challenges and the approaches they imply. For example, a lack of ground truth means researchers must use indirect ways to confirm the accuracy of their models, which includes comparing them to their colleagues' models, too. To even create a model also means needing to classify all the data and parameters available. This classification can be a key difference in approach because there are different ways of theorizing what the base data represents. Thus, a large amount of the research revolves around parameters: how to define them, how to discover and incorporate novel and

more complex parameters, and how or why to weigh them effectively. Likewise, experimental design approaches fluctuate based upon what kinds of parameters are used and how, even as there are constants. A network-based model, for example, will use a different experimental design than a feature-based model. One constant, however, is that testing these hypothesized models means having a real-world dataset, and the needs of studies (for efficient data collection and for data that has all the metrics being studied) make certain datasets reappear across the different sources.

Findings can vary greatly, but in many ways they are similar due to the shared aspects of goals and approaches. Namely, all findings connect to certain aspects, even as they vary in the exact conclusions. All sources find a conclusion relating to shared domains, things like considering active versus influential bloggers, factoring temporal aspects, replacing the absent ground truth with alternatives in models and baselines, evaluating data using established mathematical methods like correlations, and placing import of certain parameters over others. Yet, while all these domains remain implicit in findings, researchers will naturally come to different findings because of the differences in their approach. This combination—the shared but the different—hints at the value each study brings. In short, comparing each study shows us what salient findings cross all studies, and identifying their differences shows what is novel in each study and its findings.

References

Agarwal, N., Liu, H., Tang, L., & Yu, P. S. (2008, February). Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 207-218).

Agarwal, N., Liu, H., Tang, L., & Yu, P. S. (2012). Modeling blogger influence in a community. *Social Network Analysis and Mining*, *2*, 139-162.

Akritidis, L., Katsaros, D., & Bozanis, P. (2009, September). Identifying influential bloggers: Time does matter. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 76-83). IEEE.

Akritidis, L., Katsaros, D., & Bozanis, P. (2011). Identifying the productive and influential bloggers in a community. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *41*(5), 759-764.

Aziz, M., & Rafi, M. (2010, December). Identifying influential bloggers using blogs semantics. In *Proceedings of the 8th International Conference on Frontiers of Information Technology* (pp. 1-6).

Ishfaq, U., Khan, H. U., & Iqbal, K. (2017). Identifying the influential bloggers: a modular approach based on sentiment analysis. *Journal of Web Engineering*, 505-523.

Kayes, I., Qian, X., Skvoretz, J., & Iamnitchi, A. (2012). How influential are you: Detecting influential bloggers in a blogging community. In *Social Informatics: 4th International Conference, SocInfo 2012, Lausanne, Switzerland, December 5-7, 2012. Proceedings 4* (pp. 29-42). Springer Berlin Heidelberg.

Khan, H. U., Daud, A., & Malik, T. A. (2015). MIIB: A Metric to identify top influential bloggers in a community. *PloS one*, *10*(9), e0138359.

Khan, H. U., Daud, A., Ishfaq, U., Amjad, T., Aljohani, N., Abbasi, R. A., & Alowibdi, J. S. (2017). Modelling to identify influential bloggers in the blogosphere: A survey. *Computers in Human Behavior*, *68*, 64-82.

Moh, T. S., & Shola, S. P. (2013, October). New factors for identifying influential bloggers. In *2013 IEEE International Conference on Big Data* (pp. 18-27). IEEE.