Identifying Influential Bloggers Time Does Matter

Leonidas Akritidis
Dimitrios Katsaros
Panayiotis Bozanis

WI/IAT 2009, September 15-18, Milan, Italy

The Evolution of Web 2.0

- Massive transition in the applications and services hosted on the Web.
- The obsolete static Web sites have been replaced by numerous, novel, interactive services
- New Feature: Dynamic Content

Virtual Communities

- Web 2.0 includes virtual communities where the users share
  1. Ideas
  2. Knowledge
  3. Experiences
  4. Opinions
  5. Files (Media Content, Images, Audio, Video)

- Examples Include
  1. Blogs
  2. Forums
  3. Wikis
  4. Media Sharing Services
  5. Bookmarks Sharing Services

Blogs

- Blogs are locations on the Web where some individuals (the bloggers) express their opinions or experiences about various subjects.
- Such entries are called blog posts
- The readers submit their own comments to the original blog post.

Posts

- A post is characterized by the blogger's name and the publication date.
- We know who wrote this post and when.
- It may contain text, images, videos or sounds and links to other blog posts and Web pages.

Blogosphere

- The virtual universe that contains all blogs
- Accommodates two types of blogs:
  1. *Individual blogs*, maintained and updated by one blogger
  2. *Community blogs*, or multi-authored blogs, where several bloggers may start discussions
- We focus only on community blogs.

The Influentials

- In a physical world, people use to consult others about a variety of issues:
- Which restaurant to choose, which place to visit, which movie to watch.

- These others are *the influentials*.

- This is also valid for Blogosphere

Identifying the Influential Bloggers: Why is it important?

- The influentials help others in their decision making and their opinion is important.
- Companies can “use” them as “unofficial spokesmen”, instead of advertising their products.
- They are considered as market movers. (Don’t buy this, but this).
- They can forge political agendas (Don’t vote him).

Identification of Influential Bloggers

- It seems similar to the problem of identifying influential blog sites and authoritative Web pages.

- However, the techniques proposed for these problems cannot be applied to our case.

Existing Models

- Very few
- Not relative
- Blogosphere modeling, mining, trust/reputation, spam blog recognition, discovering and analyzing blog communities
- Relative:
- Influence Flow Model

Influence Flow Model (IFM, 1)

- It is based on four parameters

- Recognition (number of incoming links),
- Activity Generation (number of comments)
- Novelty (i.p. to number of outgoing links)
- Eloquence (i.p. to the post's length).

Influence Flow Model (IFM, 2)

- An influence score is calculated for each post. The post with the maximum influence score is used as the blogger's representative post.

$$I(p) = w(\lambda)(w_{com}\gamma_p + w_{in}\sum_{m=1}^{|\iota|} I_p(m) - w_{out}\sum_{n=1}^{|\theta|} I_p(n))$$

- w(λ) is a weight function of the post's length
- w_{com} regulates the contribution of the number of comments γ(p).
- w_{in} and w_{out} adjust the contribution of the incoming and outgoing influence.

IFM Drawbacks

- Isolating a single post is simplistic.
- A blogger may have written only a handful of influential posts and numerous others of low quality. Productivity is overlooked.
- It depends on user defined weights. Changing the values of the weights leads to alternative rankings.
- It ignores a very important factor: Time.
- It uses demanding and unstable recursive definitions

Measuring a Blogger's Influence

- Number of Posts (Productivity)
- Age of the Posts
- Number of Incoming Links
- Age of the Incoming Links
- Number of Comments
- We argue that the outgoing Links weaken a post's influence.

MEIBI Scores

- Metric for Identifying a Blogger's Influence
- Assigns a score to the i^{th} post of the j^{th} blogger

$$S_j^m(i) = \gamma(|C(i)| + 1)(\Delta TP_j(i) + 1)^{-\delta}|R_j(i)|$$

- $\Delta TP_j(i)$: time interval (in days) between current time and the date that the post i was submitted.
- $R_j(i)$: posts referring the i^{th} post of blogger j.
- C(i): the set of comments to post i of blogger j
- $\gamma=4$, $\delta=1$

MEIBI Definition

- A blogger j has MEIBI index equal to m, if m of his/her BP(j) posts get a score $S_j^m(i) \geq m$ each and the rest BP(j) − m posts get a score $S_j^m(i) \leq m$
- This definition awards both influence and productivity

- A blogger will be influential if s/he has posted several influential posts recently.

Motivations

- An old post may still be influential.
- How could we deduce this?
- We examine the age of the incoming links
- If a post is not cited anymore, it is an indication that it negotiates outdated topics
- On the other hand, if an old post continues to be linked presently, then it probably contains influential material.

MEIBIX Scores

- We assign to each incoming link of a post a weight depending on the link's age.

$$S_j^x(i) = \gamma(|C(i)| + 1) \sum_{\forall x \in R_j(i)} (\Delta TP(x) + 1)^{-\delta}$$

- ΔTP(x): time interval between current time and the date that the post x was submitted.

MEIBIX Definitions

- A blogger j has MEIBIX index equal to x, if x of his/her BP(j) posts get a score $S_j^x(i) \geq x$ each, and the rest BP(j)−x posts get a score $S_j^x(i) \leq x$

Experiments: Dataset

- Millions of blog sites exist
- It is essential to detect an active blog community that provides

1. Blogger Identification
2. Date and time of posts
3. Number of comments
4. Number of outgoing links.

Data Characteristics

- The Unofficial Apple Weblog (TUAW) meets all these requirements.
- Crawled in the first week of Dec. 2008.
- 160,000 pages.
- 17,831 blog posts.
- 51 unique bloggers.
- 269,449 comments.
- 5 years of blogging activity.

Inlinks: Age

- Posts get old very quickly
- The majority of links come within a few hours after the post's submission

Age	Inlinks	Percentage
0 days	26346	49,2%
1 day	13470	25,1%
between 1 and 7 days	6653	12,4%
between 7 and 30 days	2406	4,5%
between 30 and 60 days	928	1,7%
between 60 and 365 days	2523	4,7%
over 365 days	1249	2,3%
Total	53575	99,9%

Plain Methods for Bloggers Ranking

- Ranking by blogging activity
- S. Mc Nulty is the most active blogger
- He has been inactive in the last 5 months

	Bloggers	N	First	Last
1	S. McNulty	3037	06/01/2005	31/07/2008
2	D. Caolo	2242	07/06/2005	04/12/2008
3	D. Chartier	1835	26/08/2005	30/08/2007
4	E. Sadun	1560	09/11/2006	26/09/2008
5	C.K. Sample III	1057	01/03/2005	05/06/2006
6	M. Lu	1043	13/12/2006	04/12/2008
7	L. Duncan	954	19/09/2004	23/01/2007
8	C. Bohon	793	24/02/2004	04/12/2008
9	M. Rose	793	29/11/2006	05/12/2008
10	M. Schramm	648	07/06/2007	04/12/2008

Plain Methods for Bloggers Ranking

- Ranking by H-Index
- E. Sadun is the most influential blogger
- She has been inactive in the last 3 months

	Bloggers	h	Posts	Cited	Inlinks
1	E. Sadun	31	1560	489	5759
2	C. Bohon	29	793	676	9439
3	M. Schramm	25	648	339	4322
4	R. Palmer	25	354	354	4809
5	M. Rose	24	793	364	4222
6	D. Caolo	23	2242	459	4907
7	M. Lu	23	1043	397	4282
8	S. McNulty	23	3037	334	3212
9	B. Terpstra	22	226	223	3013
10	C. Warren	22	133	112	1605

MEIBI Rankings

- MEIBI considers Bohon as the most influential (793 posts, 676 cited posts, 9,439 inlinks and 14,745 comments)
- Rose is the 5th, better than Warren

	Bloggers	m	C_j
1	C. Bohon	49	14745
2	R. Palmer	46	9916
3	S. Sande	36	7246
4	E. Sadun	34	32432
5	M. Rose	30	13499
6	M. Schramm	30	12838
7	C. Warren	28	4857
8	D. Caolo	27	27985
9	M. Lu	25	17966
10	B. Terpstra	17	3770

MEIBIX vs Plain Methods

- The top four bloggers are the same.
- On the other hand, MEIBIX considers Warren to be more influential than Rose.

	Bloggers	x
1	C. Bohon	48
2	R. Palmer	47
3	S. Sande	37
4	E. Sadun	33
5	C. Warren	30
6	M. Rose	29
7	M. Schramm	27
8	M. Lu	26
9	D. Caolo	25
10	B. Terpstra	15

Comparison

- Rose: 793 posts, 364 cited posts, 4222 (5.3 per post) incoming links and 13499 comments (17 per post)
- Warren: 133 posts, 112 cited posts, 1605 incoming links (12 per post) and 4857 comments (36.5 per post).

Conclusion

- Rose has published more posts and received more incoming links and comments.
- But Warren's posts are more attractive.
- Therefore, MEIBI is more sensitive to the overall performance of a blogger (productive bloggers)
- And MEIBIX awards bloggers that publish more influential posts.

Rankings in limited time windows

- We have tested our methods by only considering the posts published only the previous month (November 2008).

	Bloggers	N	Inlinks	C_j
1	C. Bohon	47	508	556
2	R. Palmer	42	339	491
3	S. Sande	34	354	177
4	M. Schramm	29	203	166
5	D. Caolo	20	163	178
6	M. Rose	19	138	154
7	B. Terpstra	15	103	87
8	C. Warren	8	80	331
9	M. Lu	8	71	248
10	V. Agreda	5	30	42

	Blogger
1	C. Bohon
2	R. Palmer
3	M. Lu
4	C. Warren
5	D. Caolo
6	C. Ullrich
7	S. Sande
8	M. Rose
9	V. Agreda
10	Jason Clarke

	Blogger	m
1	C. Bohon	26
2	R. Palmer	20
3	S. Sande	20
4	D. Caolo	17
5	M. Schramm	16
6	M. Rose	13
7	M. Lu	8
8	B. Terpstra	7
9	C. Warren	7
10	V. Agreda	4

	Blogger	x
1	C. Bohon	27
2	S. Sande	20
3	R. Palmer	19
4	D. Caolo	18
5	M. Schramm	16
6	M. Rose	13
7	M. Lu	8
8	B. Terpstra	7
9	C. Warren	7
10	V. Agreda	4

Table IX
BLOGGERS RANKING ACCORDING TO: TUAW (LEFT). INFLUENCE-FLOW MODEL (CENTER). MEIBI AND MEIBIX (RIGHT).

Comparison

- The IFM method positions M. Lu in the 3rd place, higher than C.Warren and D.Caolo.
- But for that specific month, Warren and Caolo have published more posts.
- Moreover, their posts received more inlinks and comments. Hence, their posts were more influential than Lu's.
- MEIBI and MEIBIX produce fairer rankings

Blogging behavior over 2008: MEIBI

- We have also studied the behavior of the TUAW bloggers over 2008.
- Our models allow the observation of the ranking fluctuation.

	Jan 2008	Feb 2008	Mar 2008	Apr 2008	May 2008	Jun 2008	Jul 2008	Aug 2008	Sep 2008	Oct 2008	Nov 2008
Erica Sadun	1	2	1	2	1	4	3	4	2	-	-
Scott McNulty	2	10	8	6	6	3	4	-	-	-	-
Cory Bohon	3	1	2	1	2	1	2	1	3	2	1
Dave Caolo	4	8	5	3	5	5	6	5	6	7	4
Mike Schramm	5	4	4	9	9	8	7	6	5	5	5
Brett Terpstra	6	5	7	7	8	-	-	7	8	9	8
Christina Warren	7	6	-	8	-	7	9	-	7	8	9
Mat Lu	8	3	6	4	3	6	8	8	9	6	7
Michael Rose	9	7	3	5	-	-	-	9	10	3	6
Nik Fletcher	10	9	9	10	-	-	-	-	-	-	-
Chris Ulrich	-	-	10	-	-	-	-	-	-	-	-
Robert Palmer	-	-	-	-	4	2	1	2	1	1	2
Steven Sande	-	-	-	-	7	9	5	3	4	4	3
Joshua Eliis	-	-	-	-	10	10	-	-	-	-	-
Gilles Turnbull	-	-	-	-	-	-	10	10	-	-	-
Victor Agreda, Jr.	-	-	-	-	-	-	-	-	-	10	10

Blogging behavior over 2008: MEIBIX

- We see that E. Sadun was among the most influential bloggers two months ago, but she is currently inactive.
- From the moment R. Palmer became active, he is among the most influential.
- We can not say the same for S. Sande.

	Jan 2008	Feb 2008	Mar 2008	Apr 2008	May 2008	Jun 2008	Jul 2008	Aug 2008	Sep 2008	Oct 2008	Nov 2008
Erica Sadun	1	2	1	3	1	4	3	4	2	-	-
Scott McNulty	2	-	8	6	5	3	4	-	-	-	-
Cory Bohon	3	1	2	1	2	1	2	1	3	2	1
Dave Caolo	4	7	4	2	7	5	6	5	7	7	4
Brett Terpstra	5	5	7	7	6	-	-	7	9	9	7
Christina	6	6	-	8	-	6	9	-	6	8	9
Mat Lu	7	4	5	5	3	7	7	8	10	6	8
Michael Rose	8	8	3	4	-	-	-	9	8	3	6
Mike Schramm	9	3	6	9	8	8	8	6	4	5	5
Nik Fletcher	10	9	9	10	-	-	-	-	-	-	-
Chris Urlich	-	10	10	-	-	-	-	-	-	-	-
Robert Palmer	-	-	-	-	4	2	1	2	1	1	2
Steven Sande	-	-	-	-	9	9	5	3	5	4	3
Joshua Ellis	-	-	-	-	10	10	-	-	-	-	-
Giles Turnbull	-	-	-	-	-	-	10	10	-	-	-
Victor Agreda	-	-	-	-	-	-	-	-	-	10	10

Conclusions

- We have detected and studied the problem of identifying influential bloggers in a community.
- We proposed two novel measures for that.
- For the first time, we introduce temporal aspects to the identification of the influentials.
- The two measures also award productivity.

Thank you

Any Questions?