Evaluating the Authority in a Weblog Community
Evaluando la Autoridad en una Comunidad Weblog

Carlos Alberto Ochoa Ortiz Zezzatti[1], José Alberto Hernández Aguilar[2], Sául González Campos[1], Arnulfo Castro Váquez[1] and Julio Cesar Ponce Gallegos[3]
[1]Instituto de Ingeniería y Tecnología (Departamento de Ingeniería Eléctrica y Computación); UACJ
megamax8@hotmail.com, jsaulg@prodigy.net.mx, arncastr@uacj.mx
[2]CIICAp, Universidad Autónoma del Estado de Morelos
jose_hernandez@uaem.mx
[3]Laboratorio de Inteligencia Artificial, Universidad Autónoma de Aguascalientes
julk_cpg@hotmail.com

Article received on October 31, 2007; accepted on March 17, 2008

Abstract.
The weblog medium while fundamentally is an innovation in personal publishing has also come to engender a new form of social interaction on the web: a massively distributed but completely connected conversation covering every imaginable topic of interest. A product of this ongoing communication is the set of hyperlinks made between weblogs in the exchange of dialog, a form of social acknowledgement on the part of authors. The purpose of this paper is to understand the social implications of linking in the community, drawing from the hyperlink citations collected by the Blogdex project over three years. Social network analysis is employed to describe the resulting social structure, and two measures of authority are explored: popularity, as measured by webloggers'public affiliations and influence measured by citation of each others writing. These metrics are evaluated with respect to each other and with the authority conferred by references in the popular press.
Keywords: Web log; social nets; social network analysis; permalink

Resumen.
El medio de la Weblog mientras que fundamentalmente es una innovación en la publicación personal ha también llevado a producir una nueva forma de interacción social en la Web: una conversación distribuida masivamente pero completamente conectada cubriendo cada tema imaginable de interés. Un producto de esta comunicación continua es el conjunto de hiperenlaces hechos entre weblogs en el intercambio de diálogos, una forma de reconocimiento social a los autores. El propósito de este documento es entender las implicaciones sociales de los enlaces en la comunidad, ilustrados por las citas de los hiperenlaces almacenadas en el proyecto Blogdex a lo largo de tres años. El análisis de redes sociales es utilizado para describir la estructura social resultante, y dos medidas de la autoridad son exploradas: la popularidad, medida por las afiliaciones públicas de los webloggers y la influencia medida por las citas en cada uno de los escritos de otros. Estas métricas son evaluadas una con respecto a la otra y con la autoridad conferida por los árbitros en la prensa popular.
Palabras clave: Weblog; redes sociales; análisis de una red social; permalink

1 Introduction

The medium of weblogging differs very little from other forms of online publishing which have constituted the web since its beginnings. During its infancy, only a handful of authors were writing daily to websites identified as weblogs, but undoubtedly there were many thousands of others who updated their personal homepages nearly as frequently and in a similar writing style. What distinguishes weblogging from previous web media is the extent to which it is social, and one can say that the medium came into existence when the set of web journal writers recognized themselves as a community.

In the early days, there were only a handful of individuals who practiced the form, but with the addition of simple, personal publishing tools the community began an exponential growth that persists today. What was once a small family has matured into a burgeoning nation of millions including immense sub-communities around tools such as LiveJournal and DiaryLand. While some of these webloggers identify with the progenitors of the medium, others feel that their practice is distinct from that form. Regardless of affiliation, the nation of weblogging exists as such because every individual who takes part is connected to all others through the social ties of readership [8].

Every informal social system has its own order, constituted by the attribution of friendship, trust, and admiration between members. These various forms of social association give rise to higher-level organization, wherein individuals take on informal roles, such as opinion leadership, gatekeeper or maven. Within the weblog community, these positions are sought after by many authors, as they convey a sense of authority that increases readership and ties with other webloggers.

Related Work

Acording [16] authority is the number of blogs linking to a website in the last six months. The higher the number, the more Authority the blog has. It is important to note that is measured the number of blogs, rather than the number of links. So, if a blog links to a blog many times, it still only count as +1 toward an authority. Of course, new links mean the +1 will last another 180 days.

Attempts to mathematically determine which blogs are the most influential are currently characterized by the shortcoming of just counting blogs and links. This methodology is not unlike asking “who is the most influential scientist in the world” and then counting citations to determine the answer: neither process measures influence within genres, such as politics, cooking or technology [17]. Besides considering just counting links, we will explore the weblog’s authority by means of two additional measures: popularity and influence.

Popularity. Measured by the webloggers’public affiliations.

Influence. One way to understand the development on Blogosphere is to find influential blog sites. There are many non-influential blog sites which form the "the long tail". Regardless of a blog site being influential or not, there are influential bloggers. Active bloggers are not necessarily influential. Influential bloggers can impact fellow bloggers in various ways. An influential blogger have influential posts and collectable social gestures (statistics) like citations, comments, outgoing links and goodness of blog post [15]:

Recognition: Citations (incoming links). An influential blog post is recognized by many. The more influential the referring posts are, the more influential the referred post becomes.

Activity Generation: Volume of discussion (comments). Amount of discussion initiated by a blog post can be measured by the comments it receives. Large number of comments indicates that the blog post affects many such that they care to write comments, hence influential.

Novelty: Referring to (outgoing links). Novel ideas exert more influence. Large number of outlinks suggests that the blog post refers to several other blog posts, hence less novel.

Eloquence: “goodness” of a blog post (length). An influential is often eloquent. Given the informal nature of Blogosphere, there is no incentive for a blogger to write a lengthy piece that bores the readers. Hence, a long post often suggests some necessity of doing so.

Another way to measure influence is by means of an influence graph. An influence graph is a weighted, directed graph with edge weights indicating how much influence a particular source node has on its destination. Starting with the influence graph we aim to identify a set of nodes to target a piece of information such that it causes a large number of bloggers to be influenced by the idea [18].

Problem at hands

This paper presents an exploration of the concept of authority as it is manifested in the community of webloggers. For this purpose we use collected data from the Blogdex aggregator on the referential information contained within weblogers for three years, namely the hypertext links contained within webloggers’writing. Despite this system is out of line since last may 2006, we decide to use the data we previously collected because the importance of this site and its influence on modern sites like FlickAir, MetroFlog and MySpace, among others. We use Social Network Analysis and text mining to describe the resulting social structure and the authority distribution.

The contribution of this paper is the proposal on how to measure the authority based not just on links counting but on the popularity and influence analysis on this kind of social networks by means of Social Network Analysis.

2 Background

Social network analysis (SNA) is a discipline of social science that seeks to explain social phenomena through a structural interpretation of human interaction both as a theory and a methodology [14]. SNA assumes a basic graph representation where individuals (actors) are characterized by nodes, and the relationships (ties) they form with each other are edges between these nodes. This graph may be undirected, assuming that all social relationships are reciprocal, or directed, where each interaction describes a one-way association between two people. The degree of any node is defined as the number of associates that node has; in the case of undirected graphs, the degree is separated into in-degree (links in) and out-degree (links out).

Social scientists have characterized power as an actor's ability to control resources and information within the network, typically by exerting some type of structural advantage over other actors. Katz and Lazarsfeld made the observation that influence is controlled by a two-step flow of communication wherein opinion trickles up to opinion leaders and then back down to the rest of the population [6]. Social network showing that innovations, rumors, and beliefs tend to move from those marginal in a network, to the central figures, and back to the rest of the population [9,10, 11, 12, 13].

Opinion leaders are typically observed by their centrality to a given network, or by their ability to exercise large portions of the population in question by controlling the flow of information [4,5]. Freeman has described centrality in three different measures: degree of centrality, or the total number of ties an actor has, between ness centrality, or the probability that an individual lies on a path between any two nodes in the network, and closeness centrality, the extent to which an actor is close to all other actors [3]. Since between's and closeness centrality require a complete description of the network and considerable computational resources for large data sets, degree centrality is typically used as simple and efficient means of calculating authority.

Network analysis is well suited for the study of weblogs as many of the social relationships between weblog authors are explicitly stated in the form of hypertext links. Webloggers have posted their own interpretation of popularity and influence based on the number of links a weblog has in various link aggregation systems. Many webloggers use the total number of links to their site to evaluate the effectiveness of their writing.

A recent debate that has risen quite a bit of attention among webloggers is related to the distribution of these links within the community; Clay Shirky wrote a piece documenting the fact that a small group of webloggers had an enormous number of links to their site while the great majority only had a few [11]. This distribution, he claimed, followed a power law distribution, a widely observed phenomenon popularized recently in [1]. Shirky assumed this model to claim that within the weblog ecosystem, the "rich get richer", and that the longer one has been an author; the more central they will be.

Defining Weblog Social Ties

Weblogs are a massively decentralized conversation where millions of authors write for their own audience; the conversation arises as webloggers read each other and are influenced by each others' thoughts. A number of distinct subtypes of links have emerged within the medium, each one conveying a slightly different kind of social information:

Blogrolls

Nearly every weblog contains a list of other weblogs that the author reads regularly termed the blogroll. This form evolved early in the development of the medium both as a type of social acknowledgement and as a navigational tool for readers to find other authors with similar interests. In some hosted services, such as LiveJournal and Xanga, the blogroll is a core part of the interaction, allowing users to be notified when their friends make a post or even to create a group dialog represented by the sum of the group's individual weblogs.

Permalinks
Weblogs are comprised of many individual entries, each covering a different interest or line of thinking. During the development of the first weblogging systems, it became apparent that it would be necessary to refer to specific posts instead of an entire weblog [2]; this feature allowed authors to have a sort of distributed conversation, where one post can respond to another on an entirely different weblog. These entry reference points are called permalinks and they are a core element of nearly every weblog today.

Comments
The most basic form of weblog social interaction is the comment, a reader-contributed reply to a specific post within the site. Comment systems are usually implemented as a chronologically ordered set of responses, much like web bulletin board systems.

Trackbacks.
A recent feature of weblog tools is the trackback, an automatic communication that occurs when one weblog references another. If both weblogs are enabled with trackback functionality, a reference from a post on weblog A to another post on weblog B will update the post on B to contain a back-reference to the post on A. This automated referencing system gives authors and readers an awareness of who is discussing their content outside the comments on their site.

3 Methodology

The Blogdex project was launched in 2001 as an effort to track the diffusion of information through the weblog community and formerly is out of line since 2006. The system tracked over 30,000 weblogs, updating its index when weblogs were changed, and keeping a record of each link made on a weblog along with the time the citation occurred. These links were aggregated into an index of the most rapidly diffusing content at any given point in time. These data were made available publicly on the project's homepage [7] and we obtained our database when information was available online.

4 Analysis

4.1. Data gathering and selection

To extract the social network from the database, we first normalize the URLs of known weblogs to deal with potential duplicates, removing any leading "www" string and any trailing file name:
http://www.myweblog.com/index.html→myweblog.com.

The resulting string is termed the weblog ego as it represents a unique key to all links that come from a particular site. These strings are then queried as substrings of links in the database; when a match is found, if the normalized form of the resulting URL is the same as the weblog ego, we assume this is a blogroll link. When the URL points to content other than the front page, we presume the link is a permalink. For example, if the ego in question is "myweblog.com", we would identify the following blogroll and permalinks as such:
http://www.myweblog.com/ → blogroll
http:/myweblog.com/archives/001385.html→permalink.

The social network is then represented by recording the weblog the link occurred on, the weblog the link pointed to, and the type of link (blogroll or permalink). Since one weblog can link to many permalinks on another given weblog, we associate a weight score with these links, the value of which is simply the number of permalinks from one weblog to another.

Blogrolls and permalinks represent two different types of social reference. A link made on a blogroll is made explicitly as a statement of social affiliation. By placing a link to another weblog, one assumes that the author either endorses that weblog, wishes to promote it, or claims to read it on regular basis. To achieve this comparison, we will examine the top 1,000 weblogs by degree for both sets of network data, qualitatively analyzing the top weblogs for each ranking and looking quantitatively at the distribution of authority across these sites.

4.2. Results

The network data collected by Blogdex contained 27,976 weblogs that have at least one inbound or outbound tie. The blogroll network data consists of 116,234 ties between these weblogs while the permalink data contains 285,970 ties. The higher density of permalink ties can be attributed to the fact that these relations accrue over time, and during the process of writing a weblog, while blogroll links require an explicit effort.

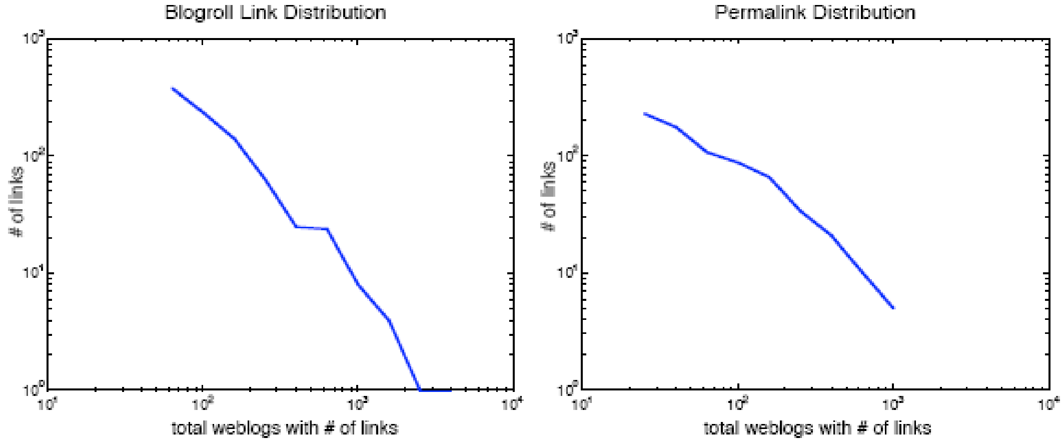


Fig. 1 & 2. Blogroll and Permalink distributions

In ranking these sites, the first observation is that most of the top sites are standard weblogs per se, they are weblog-like tools supported by communities of authors. Metafilter, Slashdot, Plastic, Fark and others depend upon tens of thousands of people for their content, while the rest of the list consists of sites operated by one or a small group of people. We can think of these systems as playing a role that a single human cannot, i.e. maintaining social ties with thousands of individuals. These sites play a crucial role in connecting large parts of the weblog network, resonating the important information that is diffusing through the community. Figures 1 and 2 show the distribution of rank across the weblog social network for both blogroll links and permalinks. The first observation from these figures is that the slope of the blogroll distribution is slightly steeper than the permalinks, suggesting that the falloff for authority in blogrolls is quicker, leaving a bigger separation between those at the tail of the curve.

Table 1. Top authoritative sites by Blogroll and Permalink degree

Rank	Blogroll Degree Rank		Permalink Degree Rank	
1	2581	metafilter.com	1322	boingboing.net
2	2434	slashdot.org	1270	diveintomark.org
3	2146	boingboing.net	1096	metafilter.com
4	1825	kottke.org	1073	slashdot.org
5	1604	instapundit.com	982	kottke.org
6	1527	scripting.com	976	weblog.siliconvalley.com/column/dangillmor
7	1307	evhead.com	956	instapundit.com
8	1220	andrewsullivan.com	828	andrewsullivan.com
9	1062	memepool.com	827	themorningnews.org
10	1007	doc.weblogs.com	826	rathergood.com
11	977	megnut.com	819	textism.com
12	961	littlegreenfootballs.com/weblog	683	denbeste.nu
13	899	diveintomark.org	626	doc.weblogs.com
14	880	littleyellowdifferent.com	625	asmallvictory.net
15	848	textism.com	582	rightwingnews.com
16	846	rebeccablood.net	577	microcontentnews.com
17	758	plasticbag.org	568	joi.ito.com
18	737	dashes.com/anil	560	buzzmachine.com
19	719	ftrain.com	553	waxy.org
20	714	plastic.com	522	a.wholelottanothing.org

Table 1 shows the top 20 sites by degree for both the blogroll and permalink network data sets. While some sites, such as Kottke, Boingboing, Andrewsullivan and Instapundit maintain high rank in both lists, as the list continues it becomes increasingly divergent. Many of the earlier "A-List" weblogs, such as Rebecca Blood, Scripting News, Megnut and LittleYellowDifferent do not place in the top 20 for permalinks, suggesting that while their names are recognized and placed on many blogrolls, they are not writing content as widely influential as those with high permalink rank. The power law distribution observed there contains many authors whose weblogs are half as old as those at the top of the blogrolls.

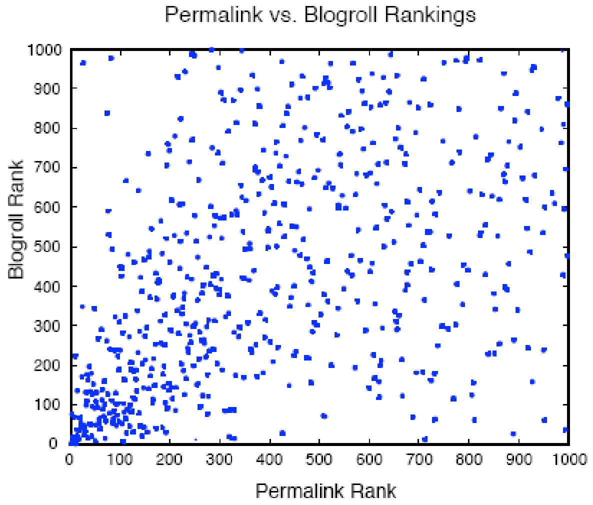
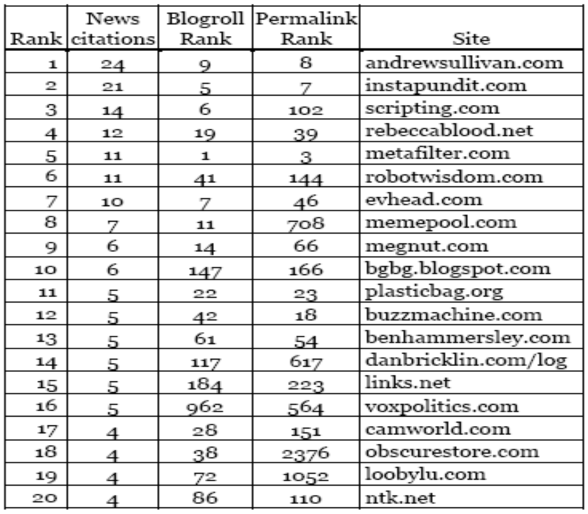


Fig. 3. Permalink and blogroll rank differential

Figure 3 demonstrates the differential between these data sets by plotting the rank in permalink versus the rank in blogrool. While the highest ranked data points tend to cluster around similar ranks, as soon as the rank passes 100 the correlation becomes much less apparent. This is further evidence that age is not the only factor in determining rank; otherwise these two data sets would be more tightly clustered around the line with slope 1. The fact that these two measures are not closely related implies that authority as measured by popularity cannot be interpreted as authority of influence.

Our queries made to Lexis Nexis returned 4,728 articles from both magazines and newspapers, of which 310 contained at least one known weblog URL. These documents yielded 545 total weblog URLs representing 212 unique sites. The twenty most cited weblogs are listed in Table 2.

Table 2. News citation rank



Rank	News citations	Blogroll Rank	Permalink Rank	Site
1	24	9	8	andrewsullivan.com
2	21	5	7	instapundit.com
3	14	6	102	scripting.com
4	12	19	39	rebeccablood.net
5	11	1	3	metafilter.com
6	11	41	144	robotwisdom.com
7	10	7	46	evhead.com
8	7	11	708	memepool.com
9	6	14	66	megnut.com
10	6	147	166	bgbg.blogspot.com
11	5	22	23	plasticbag.org
12	5	42	18	buzzmachine.com
13	5	61	54	benhammersley.com
14	5	117	617	danbricklin.com/log
15	5	184	223	links.net
16	5	962	564	voxpolitics.com
17	4	28	151	camworld.com
18	4	38	2376	obscurestore.com
19	4	72	1052	loobylu.com
20	4	86	110	ntk.net

5 Discussion

The websites collected by Blogdex were originally culled from lists of weblogs available at the time of its creation, but has since become an opt-in service for any weblogs who wish to participate. One caveat of the system is that the data set includes a selection bias based on the individuals who choose to participate. Newer weblog aggregators such as Technorati operate on an opt-out policy that creates a much more comprehensive set. Blogdex is currently transitioning towards a similar model, but the results contained within this paper are constructed from data collected under the opt-in system.

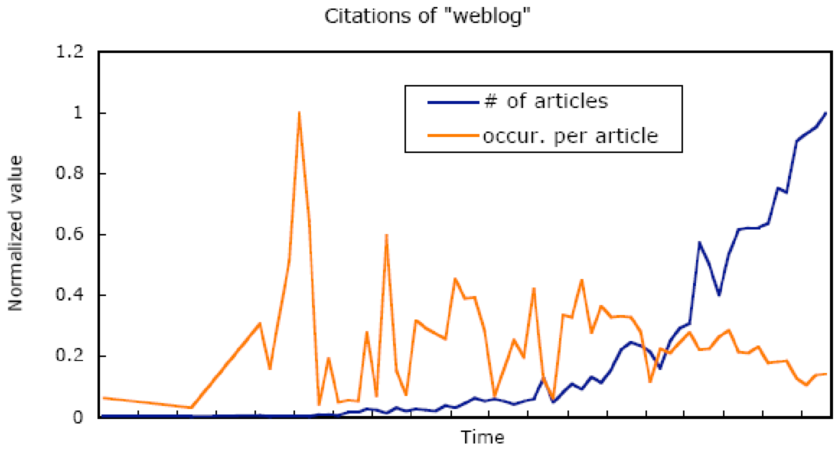


Fig. 4. Articles containing the term “weblo”, “blog”, or “web log”, and average number of usages per article related to *Timbiriche Band*

In the process of migrating the system from opt-in to opt-out, the addition of new sites to the system has been halted, resulting in a data set that is missing new weblogs.. However, since we are focusing only on the top 1,000 ranked weblog in each category, missing these newer sites should not drastically affect our results. When a more complete data set is collected, unless we have missed entire sub-networks of authors, we expect our distributions simply be shifted up by the factor of increase. Another probable explanation for the news citation rankings is that the role of weblogs in journalism is changing. Figure 4 shows the number of articles containing a weblog term versus the average number of times these terms are used per article. While the number of articles about weblogs shows an exponential growth, the number of times the term is used per article has started to wane. This is a sign that the concept of the weblog has become part of our vernacular, and as such articles about weblogging alone are probably on the decline. More recent articles are likely to be influenced more by weblogs, and less about the medium itself.

Future Work

This work is by no means complete, and could be benefited from a number of future research directions. First, this work completely ignores the dynamic element of social networks; the suspicion that blogrolls reflect a “rich get richer” scenario more than permalinks could be easily validated by examining the growth of degree for network members over time [9].

Second, not all permalinks are created equal as some weblog posts receive many orders of magnitude more traffic than others. In some cases, one post can define a weblog’s permalink rank entirely, even though this attention is quickly lost. Looking at the distribution of permalink citations for each individual weblog may allow us to renormalize the data to avoid this phenomenon. Finally, the data collected by Blogdex, while useful for the task of measuring authority, is not complete. Moving into an opt-out system and crawling weblogs at large will present a much more accurate picture of authority and influence.

6 Conclusions

The initial excitement over the weblog power law made many webloggers uncomfortable. How can a person get excited about a medium where attention is garnered by the number of weeks one has participated? Looking only at popularity by blogroll rank, it does appear that the “rich get richer”, but another assessment of authority, permalinks, might be an equally good proxy to authority and a better measure of influence. Barabási has noted that the growth of scale free networks is not only determined by the age of nodes, but also by the *node strength*, an undefined property related to a node's ability to acquire links. Permalink rank might be an accurate way of measuring node strength, and a better proxy to authority and influence at a given point in time. Many other researches related with social networks try to explain the authority in the Web [15]; In this moment Metroflog.com is analyzed in a new project for the authors of this paper with the intention that explain the ontology developed for the users of this social net when put a new image and the comments posted for this.

References

1. **Barabási, Albert-László. (2002)** "Linked: The new science of networks" Cambridge, MA: Perseus Publishing.
2. **Dash, Anil. (2004)** Interview with Paul Bausch 2003 [cited March 24, 2004]. Available from http://www.sixapart.com/log/2003/09/interview_with_.shtml.
3. **Freeman, Linton C. (1978).** "Centrality in social networks conceptual clarification". Social Networks (3): 215-239.
4. **Granovetter, Mark. (1973).** "The Strength of Weak Ties". The American Journal of Sociology 78 (6): 1360-1380.
5. **Granovetter, Mark (1983).** "The Strength of Weak Ties: A Network Theory Revisited". Sociological Theory 1:201-233.
6. **Katz, Elihu and Paul F. Lazarsfeld (1955).** Personal Influence. Glencoe, IL: Free Press.
7. **Marlow, Cameron (2004).** Blogdex 2001 [cited May 2004]. Available from http://blogdex.net/.
8. **Marlow, Cameron (2002).** "Getting the Scoop: Social Networks for News Dissemination". Paper read at Sunbelt Social Network Conference XXII, at New Orleans, LA.
9. **Ochoa A. et al. (2007)** "Discover behavior of Turkish people in Orkut" 17th International Conference on Electronics, Communications and Computers; Puebla, México.
10. **Rogers, Everett M. (2003).** "Difussion of innovations". 5th ed. New York: Free Press.
11. **Shirky, Clay. (2004)** "Power laws, weblogs and inequality 2003" [cited May 15, 2004]. Available from http://www.shirky.com/writings/powerlaw_weblog.html.
12. **Valente, Thomas W. (1995).** "Network models of the diffusion of innovations, Quantitative methods in communication". Cresskill, N.J.: Hampton Press.
13. **Weimann, Gabriel. (1982).** "On the importance of marginality: One more step in the two-step flow of communication". American Sociological Review 47 (6): 764-773.
14. **Wellman, Barry. (1997)** "Structural analysis: From method and metaphor to theory and substance". In Social structures: A network approach, edited by B. Wellman and S.D. Berkowitz. Greenwich, CT: JAI Press.
15. **Agarwal, Nitin. Liu, Huan. Tang, Lei. & Yu, Phillip, S. (2008).** "Identifying the influential bloggers in a community", Web Search and Web Data Mining, Proceedings of the international conference on Web search and web data mining. Palo Alto, California, USA, pp 207-218.
16. **Technorati (2008).** "What is authority" in http://support.technorati.com/faq/topic/71 Consulted May 1 May 2008.
17. **Gill, Kathy (2004).** "How can we measure the influence of the blogosphere?" WWW2004, New York, NY in http://faculty.washington.edu/kegill/pub/www2004_blogosphere_gill.pdf May 2008.
18. **Java Akshay. Kolari, Pranam. Finin, Tim and Oates, Tim (2006).** "Modeling the Spread of Influence on the Blogosphere" WWW2006, May 22–26, 2006, Edinburgh, UK.

***Carlos Alberto Ochoa Ortiz Zezzatti** (Bs'94–Eng.Master'00–PhD'04-Postdoctoral Researcher'06 & Industrial Postdoctoral Research'08). He has 1 book, and 7 chapters in books related with AI. He has supervised seven thesis of PhD, 11 thesis of Master and 27 thesis of Bachelor. He participated in the organization of HAIS'07, HAIS'08, ENC'06, ENC'07, ENC'08 and MICAI'08. His research interests include evolutionary computation, natural processing language and Social Data Mining.*

***José Alberto Hernández Aguilar.** He finished his Doctorate thesis in 2007 at Universidad Autónoma del Estado de Morelos (UAEM). He received the MBA degree in 2003 at Universidad de las Américas (UDLA), A.C. Since 2002, he is part time professor at UDLA, México, D.F. for the IT Career, and since 2007 year, he is part time professor at the Sciences Faculty at UAEM. Areas of interest: Databases, Artificial Intelligence, Online Assessment Systems and Marketing Research.*

***Saúl González Campos** received the B.S. degree in computer engineering from the Universidad Autónoma de Cd, Juárez in 1990, and the M.S. degree in computer sciences from the University of Texas at El Paso, USA in 2003. He is currently working towards the Ph. D. degree in telematics engineering from the Universidad de Vigo, Spain. He is currently an Professor in the UACJ, Mexico. His research interests include evolutionary computation, ubiquitous computing and distributed systems.*

***Arnulfo Castro Vázquez** received the B.S. degree in mechanic engineering from the Instituto Tecnológico de Toluca, in 1992, and the M.S. degree in computer sciences from the Instituto Tecnológico de Orizaba, in 1999. He is currently working towards the Ph. D. degree in telematics engineering from the Universidad de Vigo, Spain. He is currently an Professor in the UACJ, Mexico. His research interests include information systems, data mining and artificial intelligence.*

***Julio Cesar Ponce Gallegos** received the B.S. degree in computer system engineering from the Universidad Autónoma de Aguascalientes in 2003, and the M.S. degree in computer sciences from the Universidad Autónoma de Aguascalientes in 2007. He is currently working towards the PhD. degree in evolutionary computation from the UAA. He is currently an Professor in the Universidad Autónoma de Aguascalientes. His research interests include Evolutionary Computation, and Data Mining.*