A proof of concept for automated discourse analysis in support of identification of relationship building in blogs.

Bruno Amaral & David Phillips

SYNOPSIS

To support relationship management for public relations practice there is a need for practitioners to understand how relationships are formed, sustained and enhanced.

This paper introduces experimental, proof of concept and early findings indicating that actors in online discourse, to a varying degree, cluster round commonly understood language tokens derived from commonly understood values in interactive interpersonal and group relationships. Furthermore, the analysis indicates that the more common conceptual values are in support of tokens, the more interactive and the actors are in the relationships.

This leads to a postulate that the notion of semantic tokens and values may provide empirical evidence in support of relationship management practice.

INTRODUCTION

There is considerable literature supporting a view that relationships are at the core of public relations. Without relationships key elements in much public relations practice are not possible including, and notably, awareness, trust and reputation.

In the 2008/9 recessions trust and reputation became key to economic activity and notably the lubricant for interbank and commercial lending.

For some time, the authors have considered the elements of relationships[i].

The social science view tends to acceptance that relationships exist and human evolutionary science supported by neuropsychology and psychology evidences the powerful need for humans to build, sustain and develop relationships.

Hofstede[ii] (1995) suggested that culture is the collective programming of the mind within a social group. The programming processes will involve abstract knowledge such as beliefs, values and ideologies, as well as the more specific rules and norms underlined by the abstract knowledge.

However, there is very little grounded research into how relationships are formed, sustained and developed. Notable, there is very little grounded knowledge based on discourse analysis. Swales, describes "a class of communicative events with a shared communicative purpose" (1990)[iii]. Bazerman[iv] (1994) speaks of typified texts and remarks that "by using [them] we are able to advance our own interests".

Pedelty, Keefe, and Lithere (2008) discovered that there are some tokens such as political tokens that provide a significantly higher volume of words denoting political discussion on the fan sites of political pop stars, double in volume to that of the general pop fans[v].

Our task was to provide empirical evidence should it exists in online discourse.

The high cost of assembling large corpora, content analysis and data processing, mostly because of the high labour costs involved, have mediated against significant findings and, at best, only delivered indicative postulates.

There is considerable evidence of blogger discourse as the driver behind relationship and group building.

This offers the researcher a mass corpus conservatively estimated of the order of 118 million[vi]. Furthermore, such discourse allows interdiction in the gap espoused in Scott's thesis that there 'are two narratives transcripts in the relations between powerful groups and marginalised ones, and that the private narrative (transcript) is nearly always inaccessible to outsiders and is marked by truthful statements which cannot be said publicly because of fear of consequences[vii]'. The nature of blogging is that discourse among equals is also available to the wider constituency who, in turn, have access for interactions and involvement.

In addition, the advances in internet search, hyper-linking, word frequency analysis and, most significantly, automated, Latent Semantic Analysis, now provided science with

relatively inexpensive tools that can be applied agnostically in discourse analysis. In other words the technology can deliver through quantitative analysis insights that were only available through qualitative analysis only a few years ago.

This paper describes the use and application of such technologies in blogging discourse to help identify how relationships are created and nurtured using discourse.

Unlike analysis hitherto, the research has no reliance on human selection of the corpus, content analysis or data analysis in a triangulated research process.

The findings indicate that actors, to a varying degree, cluster round commonly understood tokens derived from commonly understood conceptual values in interactive group relationships. Furthermore, the analysis indicates that the more common conceptual values are in support of tokens, the more interactive and the actors are in the relationships.

From these findings the authors postulate that relationships between actors are sustained and developed by tokens which have commonly held values.

WHAT WE ARE TRYING TO DO

In this paper we are questioning if relationships form due to the recognition of tokens with shared values.

Using blogs we can identify the networks formed as a result of the formation of relationships.

We argue that relationships online form as a result of two drivers: people seek out relationships with their content, and by interacting with content they find. In blogs, actors interact by commenting, sharing, or linking to content.

The research model examines three forms of inter-relationship: hyperlinks, semantic content and word clusters.

Hyperlinks

Hyperlink can be seen as more than just a unique resource locator. They can be regarded as a token. They are representative of the content that attaches to the link such as a web page, 'deepweb' database or computing process.

For hyperlink relationships we took into account outbound hyperlinks on the homepage and/or sidebar of a blog

Predominantly, people generally link to content of which they approve, they may also link to pages, articles of which they disapprove. The question we attempted to answer was why actors link to particular types of content.

Testing for relationships in discourse

We tested content for recognition of tokens and the way those tokens are described (which we identified as values).

Our methodology to identify such vales was both by word clustering and latent semantic analysis.

By selecting a random collection of blogs, identifying both their latent semantics and network of links we hope to demonstrate a correlation in these variables.

METHODOLOGY

Our research required harvesting both inlinks and outlinks from blogs and performing Latent Semantic Analysis and Word Density Analysis.

To perform word density analysis, we used a PHP script from SEO Book.com (http://tools.seobook.com/general/keyword-density/source.php). This program was later modified to perform batch analysis on a set of links.

In order to perform the same analysis for Latent Semantics, we develop specific software which is available at http://www.netreputation.co.uk/summariser/getconcepts.php

To harvest both inlinks and outlinks, two different pieces of software were used. Yahoo's Site Explorer (https://siteexplorer.search.yahoo.com/) and Dale Hunscher's Urlnet Python Library (http://www.southwindpress.com/urlnet/). This last piece of software harvests the outlinks from a given page or sets of pages and follows them creating a file that can be read by Pajek or GUESS.

Both are used to build network graphs, we selected Pajek, because urlnet's examples tend to use this format for output to retain data integrity.

GATHERING A SAMPLE FOR ANALYSIS

On December the 19th, we ran a google search with the following query: *and if about the site:blogspot.com*. The purpose was to find 12 random blogs to serve as our basis of analysis, the first twelve with a technorati authority above 40 composed the following list:

- http://ifitshipitshere.blogspot.com/ If It's Hip, It's Here
- http://ibloga.blogspot.com/ Infidel Blogger's Alliance
- http://jeremiahgrossman.blogspot.com/ Jeremiah Grossman
- http://althouse.blogspot.com/ Althouse
- http://norfolkblogger.blogspot.com/ Norfolk Blogger
- http://enclave-nashville.blogspot.com/ Enclave

- http://glenngreenwald.blogspot.com/ Unclaimed Territory - by Glenn Greenwald (Observation: last updated in Feb. 2007)
- http://downwithtyranny.blogspot.com/ DownWithTyranny!
- http://powerofnarrative.blogspot.com/ Once Upon a Time...
- http://gregmankiw.blogspot.com/ Greg Mankiw's Blog
- http://lefarkins.blogspot.com/ Lawyers, Guns and Money
- http://theoutfitcollective.blogspot.com/ The Outfit: A Collective of Chicago Crime Writers

Each was used as the starting point to look at the surrounding network.

The reason why we chose Technorati to focus on inlinks as a metric of analysis. This way we could make sure that the blogs selected were not isolated in the network for a proof of concept.

After gathering this sample for analysis, we retrieved their inlinks using Yahoo! site explorer and exporting them to a TSV file.

The TSV file displayed the first one thousand inlinks to each of the blogs in the sample.

Thus, we narrowed the sample down to the unique domain names identified for each set of inlinks. Since our analysis focuses only on blogs, inlinks were again narrowed down to include only wordpress.com, typepad.com, and blogspot.com domains and sub-domains.

USING KEYWORD DENSITY ANALYSIS

When using keyword density analysis of texts, commonly used words are revealed after elimination of frequently used words (stop words) . The purpose behind using such a technique was to see if there were common words in networks that helped to define the network.

USING CONTENT ANALYSIS

Latent Semantic Analysis (LSA) is a mathematical/statistical technique for extracting and representing the similarity of meaning of words and passages by analysis of large bodies of text. It uses singular value decomposition, a general form of factor analysis, to condense a very large matrix of word-by-context data into a much smaller, but still large-typically dimensional-representation (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990)[viii].

The similarity between resulting vectors for words and contexts has been shown to closely mimic human judgments of meaning similarity and human performance based on such similarity in a variety of ways. LSA significantly improves automatic information retrieval by allowing user requests to find relevant text on a desired topic even when the text contains none of the words used in the query.

For these reasons we applied LSA to identify the linguistic concepts in the corpus of each blog and then sought common concepts in other blogs in the network.

This approach would agnostically identify common concepts in networks.

The convergence of blogs by hyperlink, words cluster and or semantic concept would indicate content is common between blog entities and that such content was significant in relationships.

Should any combination of hyperlink, word or concept coincidentally identify the same blogs in the network would indicate a discursive link and thus discursive elements that are important in the relationships. Convergence of all three elements would demonstrate that common content in discourse has high significance in relationship maintenance.

THE TEST

Using URLnet, we inputted each set of inlinks to an example script:

```
# placeholderroot1.py
from urlnet.urltree import UrlTree
some_msn_melanoma_urls = (
'http://altphotoimages.blogspot.com',
(…)
'http://www.phyllispatterson.blogspot.com',)
net = UrlTree(_maxLevel=2)
success = net.BuildUrlTreeWithPlaceholderRoot(\
    rootPlaceholder="http://ifitshipitshere.blogspot.com/",\
    Urls=some_msn_melanoma_urls)
if success:
    net.WritePajekFile('ifitship-placeholderroot1', 'ifitship-placeholderroot1')
```

Figure 1 - URLnet Python Script used to gather network data

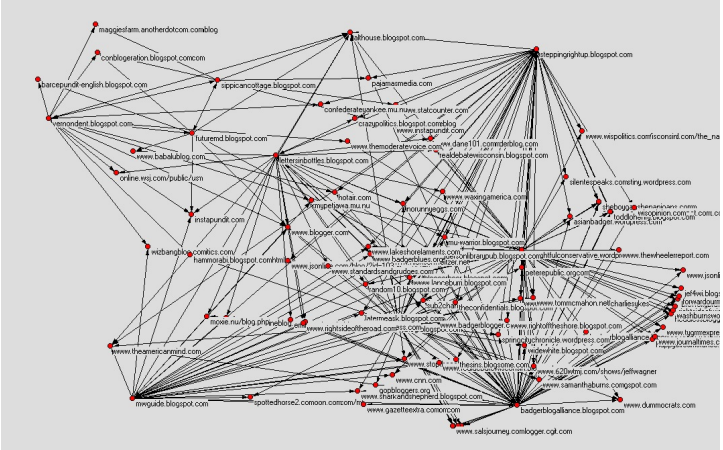
At this stage, it is important to mention that URLnet proved to be an invaluable tool, but not without it's obstacles. It did not make a distinction from links in the sidebar of a blog and those present in posts or other pages other than the homepage. Also it was not able to parse hyperlinks found in javascript code, even when that javascript's function is to output dynamic html code. It also ignored the nofollow html tag, meant to signal search engines not to follow that hyperlink.

This however did not stop us from gathering a series of network graphs. From the given list of inlinks, URLnet gathered outlinks found at each address and mapped the connections it found.

In order for the data to become readable, the output was parsed in Pajek to provide a more user friendly display. Pajek offers three possibilities to reduce a network: through the number of inputs to a vertice, the number of outputs, or both combined.

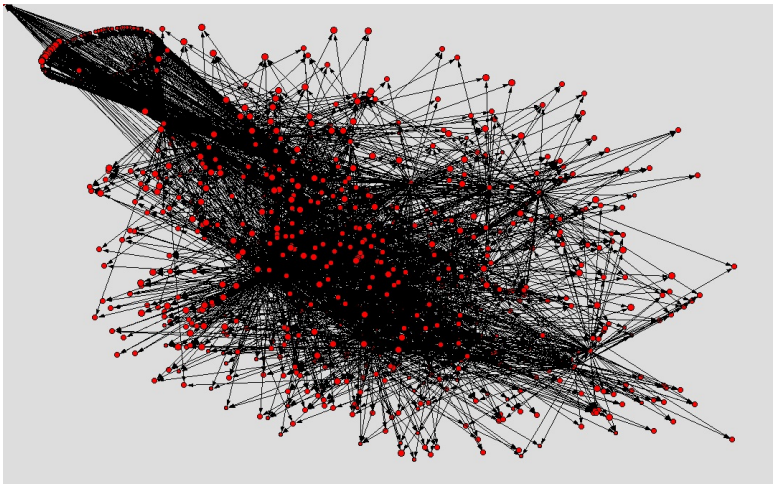
We were able to obtain several graphs using the Fruchterman-Rheingold algorithm, most of which could be read fairly well. But for larger networks this simple reduction proved to be insufficient. In this context, input and output is the equivalent to inlinks and outlinks.

As an example, some graphs could be read fairly well:



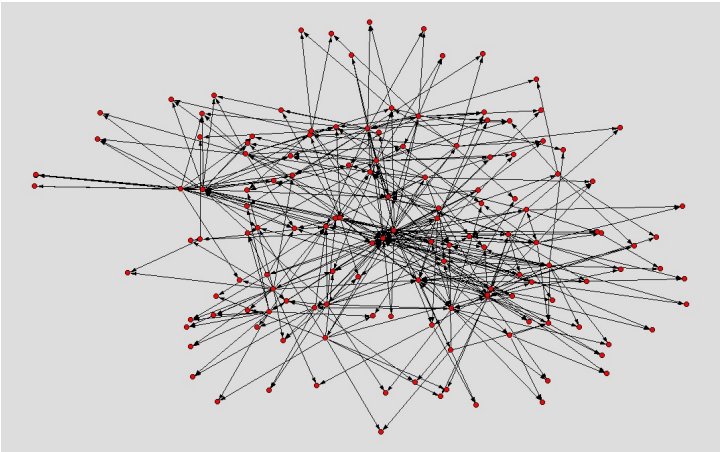
Althouse — Reduced to vertices with 2 or more inlinks.

While others still presented themselves as a complex network:



Norfolk — Reduced to vertices with 2 or more inlinks.

The blog that showed some of the closer matches between the LSA and the word density analysis was Jeremiah Grossman's. After mapping it's network using URLnet and pajek, we obtained the following graph:



Jeremiah Grossman

FINAL CONSIDERATIONS

In analysis of blogs as discourse, we identified that there was a close correlation between words used, semantic analysis and hyperlinks among bloggers. Some were more closely allied showing that where there is a more common language, the Hofstede hypothesis is born out.

At the same time, using URLnet we were able to graph and examine the way hyperlinks connect blogs in a network, seeing how disperse or close together the blogs are and identify when two blogs create hyperlinks to a third, thus triangulating more elements of the network.

Relationships are then most evident when meanings identified in texts by the LSA converge.

This offers the prospect of a proof of the nature of relationships in which common understanding of tokens and associated values are the building blocks of relationships.

Should this be proven it would provide sound evidence that alignment of values is the most powerful means for creating, sustaining and developing relationships in the domain of public relations practice.

It would be interesting to see how the influence flows through the network, do concepts flow from the center to the edges or the other way around?

The goal could be to understand which blogs quote the ones near the core of the network and which quote blogs found at the edges. Another way to look at the network would be to overlap both the graphic representation with a vector of the mains LSA categories.

The technique used to map the network could also be useful as a means to define a set of blogs to monitor, observing the dialog and mapping it for key concepts.

REFERENCES

Phillips, D (2006) Towards relationship management: Public relations at the core of organisational development. Journal of Communication Management, Volume 10, Number 2, pp. 211-226(16).

Helmond, A (2008) How Many Blogs Are There? Is Someone Still Counting? The Blog Herald http://www.blogherald.com/2008/02/11/how-many-blogs-are-there-is-someone-still-counting/ accessed Feburary 2009.

Scott, James (1990) Domination and the Arts of Resistance, New Haven: Yale University Library.

Google Semantically Related Words & Latent Semantic Indexing Technology : SEO Book.com. Available at: http://www.seobook.com/archives/000657.shtml [Accessed February 15, 2009].

Hofmann-SIGIR99.pdf. Available at: http://cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf [Accessed February 15, 2009].

Latent semantic analysis - Scholarpedia. Available at: http://www.scholarpedia.org/article/Latent_semantic_analysis [Accessed February 15, 2009].

Latent Sematic Indexing Tutorial. Available at: http://www.puffinwarellc.com/p3b.htm [Accessed January 16, 2009].

LSA @ CU Boulder. Available at: http://lsa.colorado.edu/ [Accessed February 15, 2009].

Camarinha-Matos, L. & Macedo, P., A conceptual model of value systems in collaborative networks. Journal of Intelligent Manufacturing. Available at: http://dx.doi.org/10.1007/s10845-008-0180-7 [Accessed January 16, 2009].

Grunig, J.E., 1992. Excellence in Public Relations and Communication Management 1st ed., Lawrence Erlbaum.

Hofstede, G. & Hofstede, G.J., 2004. Cultures and Organizations: Software of the Mind 2nd ed., McGraw-Hill.

Ledingham, J.A. & Bruning, S.D., 2001. Public Relations As Relationship Management,

Phillips, D., 2006. Towards relationship management: Public relations at the core of organisational development. Journal of Communication Management, 10, 211-226. Available at: http://www.ingentaconnect.com/content/mcb/jcm/2006/00000010/00000002/art00007 [Accessed January 16, 2009].

Thomas Landauer]], P.W.F. & [[Thomas Landauer]], P. W. Foltz, & D. Laham, 1998. Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259–284. Available at: http://lsa.colorado.edu/papers/dp1.LSAintro.pdf [Accessed January 14, 2009].

Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere | Berkman Center. Available at: http://cyber.law.harvard.edu/publications/2008/Mapping_Irans_Online_Public [Accessed February 15, 2009].

[i] Phillips, D (2006) Towards relationship management: Public relations at the core of organisational development. Journal of Communication Management, Volume 10, Number 2, pp. 211-226(16).

[ii] Hofstede, G. (1995) `Multilevel Research on Human Systems: Flowers, Bouquets and Gardens' , Human Systems Management 14: 207—17.
[iii] Swales, J.M. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge:
Cambridge University Press.
[iv] Bazerman, C. (1994). Systems of genres and the enactment of social intentions. In: Freedman & Medway (1994) p. 79-101.
[v] Pedelty,M., Keefe, L,. and Li, B 2008 Applying Computer Aided Content Analysis to Blogs: A Study of Politics, Popular Music, and Fan Blogging Insitute for New Media Studies http://www.inms.umn.edu/media/NMR@UMN%20-%20C%20-%20Pedelty.pdf Accessed February 2009
[vi] Helmond, A (2008) How Many Blogs Are There? Is Someone Still Counting? The Blog Herald http://www.blogherald.com/2008/02/11/how-many-blogs-are-there-is-someone-still-counting/ accessed Feburary 2009.
[vii] Scott, James (1990) *Domination and the Arts of Resistance*, New Haven:
Yale University Library.
[viii] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. A., Indexing by latent semantic analysis , Journal of the American Society for Information Science, 1990, 41(6), 391-407.