

## Unedited Interview with Dr. Edel Garcia

July 22, 2006

This is the unedited/unrevised version of a recent interview I provided to my good friend and supporter Mike Grehan for his featuring column at ClickZ. The column was tentatively titled

### The myths and maths of SEO

Mike mentioned that due to limitations in space he can only include snippets of it, but that he will be linking his column to the piece. Enjoy it.

#### Interview with Dr. Edel Garcia

- 1) Could you give us some background to your academic and commercial experience in the field of information retrieval?

I earned a PhD from ASU, Tempe, AZ in 1995. My thesis was on applied fractal geometry and chaos theory. The goal was to investigate the growth of branching patterns in physical systems like in electrodeposition of metals and whether these belong to the family of growth described by diffusion-limited aggregation models (DLA).

In DLA sticking particles are considered random walkers moving under the influence of a diffusion field. The result is a diffusion front of random walkers sticking to the active sites of a diffusion interface, which is also moving. A rich-get-richer phenomenon takes place. Multifractal clusters can be formed with variants of the model. We were able to demonstrate that in real systems an interplay between the diffusion, electric and convective fields can be established in certain cases. Electric anisotropy and flow dynamics cannot be overlooked. Thus, some fractal patterns might not belong to plain DLA.

In material sciences, mechanical and physical properties have been associated with the morphological complexity of these materials. Such patterns and diffusion interfaces are found elsewhere, in many dissimilar fields, including Web repositories. In fact, today there is an entire area of research dedicated exclusively to fractals in information retrieval. I maintain a resource page for those interested in the subject. Prior to 1989 I conducted research work in chemometrics, especially in simplex optimization processes. Thus, I guess my transition to a computer science sub discipline like IR was quite transparent.

Some times it makes me laugh when hearing from certain low-level search engine managers that non-CS scientists, like physicists or chemists, do not belong to IR. Nonsense. I call this intellectual racism. I could say the same of IRs getting into quantum mechanics. They should read "The Geometry of Information Retrieval" of Keith van Rijsbergen, one of the fathers of modern information retrieval or look at how condensed matter physicists are now modelling collections using diffusion geometries and particle statistics. Applied scientists have a solid background on things that pure CS scientists are mystified about or are trying to reinvent because of commercial or stock share pressures.

- 2) As a recognized expert and practitioner in the field of information retrieval, how and when did the SEO community become known to you.

My transition to SEO was a consequence of my transition to IR from Science. I've been involved with SEO even before some very vocal 'experts' thought about getting into the field. I even worked back in 2000 on a search engine patent for a now defunct NASDAQ company and helped to put together R&D departments on several companies before many of these folks made their own transition to SEM. Two years ago I simply got tired of so many speculations in the industry and decided to do something. That something was trying to help seos to get educated and professionalized.

- 3) What were your first impressions of the level of knowledge and understanding of IR theory and practice in the SEO community?

Pure nonsense. However, in the interest of fairness I could say the same about the wrong impression many CS scientists have about SEOs and SEMs, labelling them as a bunch of spammers.

Many IR and PhD students have developed beautiful models in controlled computer lab environments. When transferred to a real environment like the Web, these models simply do not work. To illustrate, in a commercial, noisy environment with all sort of vested interests and strategic alliances, equating literature citation to link citation or to a vote of citation importance is fallacious. Yet, many respected university colleagues bought this fallacy and others are still defending it. Like these many SEOs and marketers bought that fallacy, but motivated by vested interests.

- 4) You have a record achievement of attracting over 70,000 views to one of your posts alone over at Danny Sullivan's Search Engine Watch forums (Keywords Co-occurrence and Semantic Connectivity). We're you surprised at that level of interest and what sort of feedback did you receive?

No. There is a sector that is hungry for knowledge. Regarding trying to educate others by posting in discussion forums, I have now mixed feelings as -in my opinion- these are not the best educational environments. I prefer now to use a less noisy environment like my own site and a recently launched blog. The goal of the blog is to make visitors rethink about things they might have read before in other places. At the same time, I'm investing content, time and effort in my own web property and not in someone else site or web property. This does not mean that I will never post in other forums. It only means that I no longer see that as a priority.

- 5) Information related to SEO abounds on the web. There are hundreds and hundreds of forums and blogs dedicated to the subject. Yet the quality of information, in particular when related to information retrieval theory and techniques, seems to vary dramatically depending on the source. You coined a few phrases on your blog which put a little grin on my face recently. Blogonomies, Blogorreah and linkphilis are new terms to the industry, could you explain them 😊

These are social behaviours worth to study. I coined blogonomies as plural for blogonomy. Like blogorrhoea, blogonomy has been coined before to mean something different from what I mean.

I define a blogonomy as the dissemination of false knowledge through electronic forums and blogorrhoea when the dissemination is intentionally done for a profit or commercial interests. These behaviours are prevalent in the blogosphere, from here the "blogo-" stem.

A blogonomy can be the result of ignorance or speculations; nothing that a damage control campaign can fix to save face.

Professor Jon Klienbergh has researched the concept of 'burst' in the blogosphere. Along that line, while not a requirement, a "blogorrhea" outbreak can be observed when a blogospheric "burst" is the result of a blogonomy.

Spreading a blogonomy by means of not using a "link condom" leads to "linkphitis". Often the spreading agents are link spammers or someone posing as respected authority. Linkphitis is a condition involving links pointing to documents with corrupted knowledge like blogonomies. Thus, weight transmitted through these links can be considered infected weights. Note that here we are not talking about mere link farms or off-topic content but about the spreading of false knowledge and a special kind of burst. Most link models score weights but overlook this problem.

By a 'link condom' I mean any mechanism that will prevent the transmission of weight from corrupted links. Such mechanism is not just a mere 'no follow' attribute. So far the only 'protection' are human reviewers. Tagging is not an option since this has its own stagers (spam taggers) and magers (malicious taggers) to deal with. The problem is pervasive: how could one use human reviewers or 'link condoms' in large-scale databases that are continuously changing?

Want to hear about some blogonomies? A good example can be found in the so-called "keyword density" concept promoted by certain SEOs. And how about those 'experts' claiming that tokenization is stop word removal, that tokenization reduces terms to stems or that there is no concept for document normalization in term vector models? Wait, there is more.

How about those that use c-index calculators and tables without a clear understanding of association theory or on topic analysis? And how about the so-called sandbox effect? One thing is perceive a by-product or artefact as an aging or delay effect and something different is trying to convince others that if you use certain optimization strategies Google will place you in a sandbox.

I still have a 2005 laugh thanks to those 'experts' from SES NY and SES San Jose that suggested the use of single keyword metrics made out of dissimilar database metrics and that 'if you do this or that' Google will place your site in a sandbox. I'm putting together a resource page called **SEO Blogonomies**, which will list these and other irrational ideas.

- 6) Like myself, you have a reputation for your desire to demystify and debunk a lot of the half-baked theories, which frequently reach epidemic proportions within the SEO community. Your many tutorials have helped people new to the industry and to IR science get a firm grasp of the principles. And you've recently started a three part series introducing the math and debunking some myths of link based ranking algorithms. What are the more salient points of the tutorial that you wish to bring to the attention of the SEO community?

That without valid knowledge they have no hope. To illustrate, some times those that sell link building services quote papers about link models, not knowing that the term "rank" of a link graph is used in those articles in reference to the rank of a matrix and not in reference to any web page ranks (i.e., positioning of search results). Others even have claimed that there is a mythical Search Engine Markov Chain used by search engines to identify patterns in keywords, web pages and sites.

- 7) Keyword analysis is a vital factor regarding the success of a search marketing campaign. Your understanding of keywords and semantic relationships is way ahead of the average search marketer. What advice would you give when considering the structure of the copy on a web page?

Write for your main audience and as natural as possible. Use semantic-rich concepts and reinforce this with relevant context, but don't go overboard. Structure documents using the HTML DOM Model. Don't waste your time buying links, in link exchange programs, or with keyword density myths. Ignore blogonomies.

Unless you are limited to targeting a specific region, avoid regionalisms altogether. Let me give you two examples. English hyphenation rules are different in the USA and in the UK. Spanish translations from countries speaking the language might convey different meanings in another Spanish speaking countries. Only because you bought Spanish translation services from Mexico or other countries this does not mean that the end product is suitable for readers from other countries from Latin America. I have seen Spanish-translated press releases from certain search marketing associations that are laughable because were not written using neutral Spanish. The bottom line: write for the correct audience. If you are not targeting a specific region, use a neutral copy style.

- 8) Prior to the web (and early web) the document ranking mechanism was based specifically around the words on a page and the vector space model. Hyperlink connectivity based algorithms added another dimension to ranking which became much more visible with the emergence of Google. How much do you think statistical analysis of end user patterns and behavior is likely to play in the future?

Matching keywords by means of queries to terms embedded in text can fail to retrieve documents with relevant semantics. A system can do better by matching contexts to contexts and queries. This was one of the goals of Latent Semantic Indexing (LSI) via singular value decomposition (SVD) and of Information Spaces (IS) via principal component analysis (PCA). The problem is that these are computationally expensive for mega collections. How would you satisfy users expecting responses in less than a second with these technologies? Thus, LSI is more suitable as an auxiliary solution for representative samples or for sets that have been prescored or prefiltered rather than for raw collections consisting of billion of documents.

Instead of keyword-to-term matching, one can do better by matching contexts to contexts. Then we can use statistical analysis to preclassify documents by contexts and build clusters of contexts. Then query precached clusters rather than the original database. From the optimization side, how could SEOs make documents appealing for such systems? Here is where on topic analysis and high degree co-occurrences (like in-transit co-occurrence) plays a role.

About end user patterns and behaviours, these can be studied through time series analysis. Back in the early 90's I attended conferences on chaos in spatio-temporal dynamical systems and learned about attractor reconstruction from time series, Poincare Maps and other techniques designed for finding patterns and trends in natural phenomena evolving in time.

Similar analyses can be done with for example search volume data and large user's data that is time-dependent. Indeed, one can propose an entire offer-demand economic model from the Web where search volume (query data) is considered search **DEMAND** and document volume (search results) is search **OFFER**. Then we could apply to that model all the vast economic knowledge already available. In such offer-demand scenario I envision, search engine analytics could be used to value products and services online. This adds whole new dimensions and meanings to keyword research.