Project Narrative:
Building a Predictive Model of User Dynamics on a Social News Aggregation Site

John Doty

Table of Contents

I. Overview

A. Social News Aggregation Sites

A major issue confronting consumers of media in the information age is separating the signal from the noise. With millions of users, websites and programs submitting billions of pieces of content everyday how can one find what is relevant to one's interests? Two approaches have proven to be effective enough to garner millions of customers: intelligent search and collaborative filtering.

Companies like Google have constructed intelligent algorithms to parse and analyze the semantics of our queries (does "hot dog" mean he wants a picture of Labrador in a desert or a dubious snack associated with baseball?) and give us a list of results that should match our needs. When not searching for something specific we can still turn to forums and newsfeeds that are curated by algorithms such as Google News and Techmeme.

Collaborative filtering is another popular method for separating the wheat from the chafe on the Internet. Many services have been successful in offering different takes on this general idea. Facebook is increasingly integrating itself with content providers on the web through features such as "Like" buttons that essentially allow one to have a newsfeed of content your friends have deemed worthwhile. Other sites such as Hackernews have focused on using collaborative filtering in a small well-defined area targeting on content relative to startups and the technically inclined.

Reddit and Digg are two social news websites that allow users to submit links to content, vote on material that they find relevant and interesting, and contribute to comment threads that discuss the content. On Reddit, users gain "Karma" when their links and comments are up voted. Although outside of the Reddit universe this "Karma" currency is valueless, it is a part of a growing trend of gamification, "the use of game play mechanics for non-game applications particularly consumer-oriented web and mobile sites, in order to encourage people to adopt the applications."(wikipedia.org/wiki/gamification) This approach has proven wildly

popular. On a blog post on July 15, 2010, a Reddit administrator reported that in the previous thirty days Reddit had received eight million unique visitors and four hundred million page views and they have continued to grow since then.

B. Motivation for Project:

Reddit has the properties of complex adaptive systems. It is made up of millions of independent agents attempting to maximize their enjoyment by being presented quality, pre-vetted content or gaining Karma; parts of the Reddit system display distinctly nonlinear relationships. For example, reposting content is viewed as taboo on Reddit but happens consistently. Occasionally a user who posted a particular piece of content that received very little attention on Reddit will see a later post made by another user that is a link to the same content (a repost) receive thousands of votes and make it to the front page of Reddit. That first user will make a follow up post juxtaposing the two posts and lamenting the misappropriation of the comment that could have been theirs.

Being an avid user of Reddit and witnessing phenomenon like the above example, I have often been curious if one could model the way users interact with the site through comments and voting and if using such a model one could predict what submissions would gain attention and up votes and which would fade into the oblivion of Reddit's archives. The example of two posts linking to identical content but producing wildly different outcomes in terms of score, comments, and attention, indicates that the ranking of posts on Reddit is not a deterministic sorting based upon the content of the post but is rather the result of a complicated nonlinear system.

Aside from my interest in collaborative filtering in general and Reddit specifically I had a few technical goals when designing this independent research project that influenced its focus and the methods I used. I used this project as my first introduction to the Python programming language, the process of programming using a website's API, constructing and managing a database, and using Fuzzy Cognitive Maps to model a complex system.

II. Generating a Reddit Dataset

A. Gathering Data

Before constructing a model of Reddit it was necessary to gather raw historical data of Reddit content. Reddit has an open web API that allows third-party software to query Reddit links by URL and receive the data from that URL in JSON (Javascript Object Notation), a lightweight data-exchange format. In order to build a model of Reddit user dynamics it would be necessary to have data monitor and record the way new Reddit posts and comment threads developed over time. To this end, I built a custom scraper and database system.

Abstractly the Reddit scraper was primarily composed of three parts:

- Get New Posts

The first would perform a query to the (www.reddit.com/new) page getting the most recently posted content. It would then perform a set-wise comparison operator to determine which posts it had not yet encountered and pass those posts to the monitoring module.

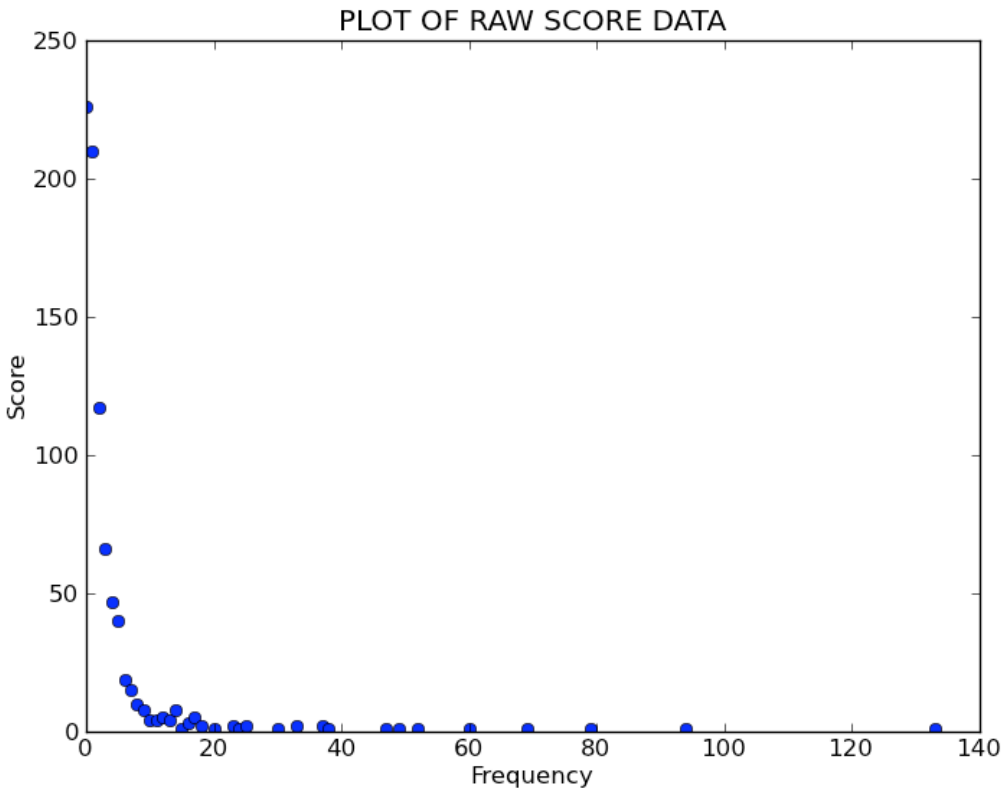
- Monitor Post

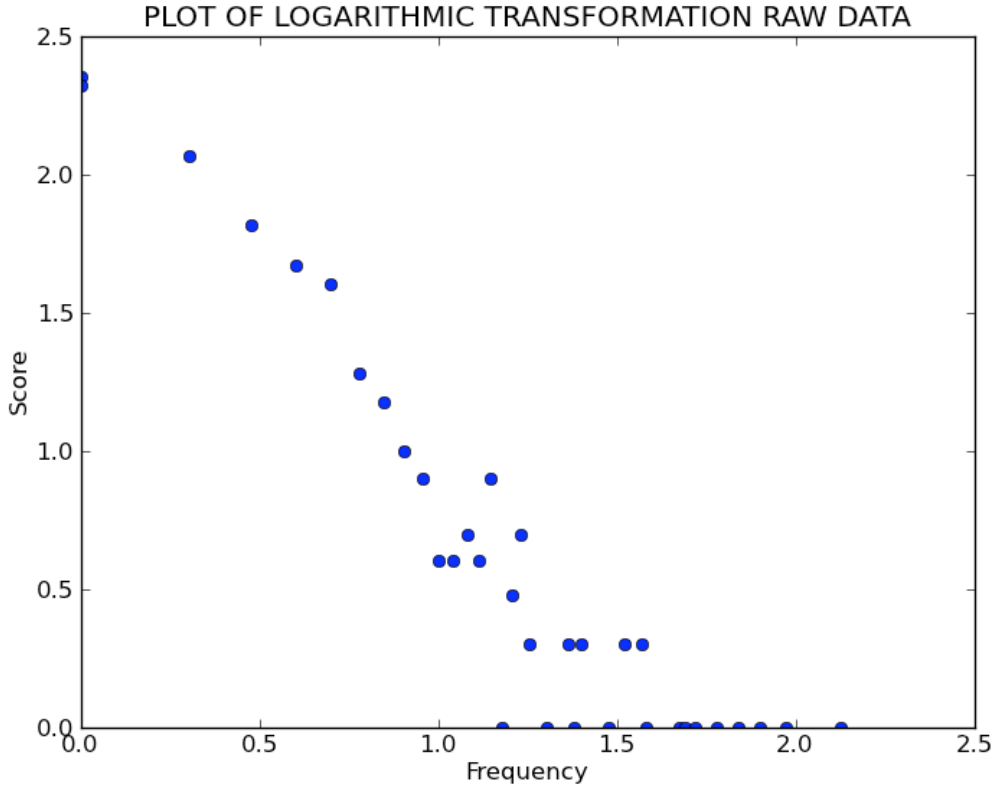
The monitoring module contains a set of posts to monitor over time. It queries Reddit for data about the state of the posts and associated comment threads it is monitoring every x minutes (where x is designated by the user of the module), creating a time series of post state data as well as incrementally backing up its data to a text file.

- Database Interface

After the scraping job is complete, the data is formatted and inserted into a CouchDB database. CouchDB is an open source document-based database that stores data in the JSON format.

The following graphs are plots of the score frequencies for a dataset comprised of 815 Reddit posts two hours after they were initially submitted.


PLOT OF RAW SCORE DATA
Score
Frequency
0
50
100
150
200
250
0
20
40
60
80
100
120
140


PLOT OF LOGARITHMIC TRANSFORMATION RAW DATA
Score
Frequency
0.0
0.5
1.0
1.5
2.0
2.5

B. Analysis of Social Voting Patterns on Digg

Similar work has been done in analyzing the voting behavior on Digg (Lerman, Ghosh). Digg allows its users to create asymmetric friend relationships, similar to "following" on twitter. Using the Friends Interface, users can track what their Digg friends are submitting and voting on. The work of Lerman et al. focused on the effect this directed friend graph has on the voting behavior on Digg. The researchers found that by looking at where the first dozen or so votes originated they could predict with a good deal of accuracy how popular the story would become as measured by total votes received. If most of the initial votes came from direct neighbors of the submitter this indicated interest within a narrow community. If however, many of initial votes came from users who were disconnected from the original poster or were a larger path length from her, the story was much more likely to be interesting to a larger audience and garner more votes.

On Reddit, there exist a similar asymmetric friend feature, but it seems to be used to a much lesser extent and thus has a less pronounced effect on voting behavior than its Digg counterpart. Thus when attempting to determine the general appeal, or 'meatiness' of a post early in its lifetime it is necessary to resort to different metrics. I hypothesized that some of the properties of the comment threads associated with a post might be early indicators of how captivating post is and its potential to garner a large number of votes and attention.

C. Comment Thread Analysis

When Reddit was initially founded it functioned purely as a post board. Users had the ability to post content and either vote posts up or down. Now every post has an associated comment thread, and this feature has become an integral part of the Reddit user experience. In the comment thread, users can either submit their comments as a direct response to the post (I call these comments 'roots')or they can submit a comment as a response to another user's comment, generating conversations trees.

There are two ways of analyzing the comment threads associated with a post.

1) The content of the comments
2) The structure of the comment threads

My approach for this project was to focus on the structure of comment threads, but I did research a few approaches to analyzing the content of comments as well as programmatically implement a Markov model that would measure the entropy of an information source based on the work of C. E. Shannon. The general question I had hoped to answer was whether there was a correlation between post score data and the novelty (entropy) of the language used in the comments associated with it. This algorithm was not used in generating data for the model discussed in this paper however.

A comment thread associated with a particular post can be represented abstractly as a set of tree structures. Each tree is made up of a root comment (defined above) and all of the comments that respond to it or any of its descendants. Using this abstraction I created several functions to analyze the graph theoretic properties of Reddit comment threads hoping to find correlations between their topological features and final score data. The following is a list of the graph theoretic properties I analyzed for Reddit comment threads:

- Ratio of stub nodes to conversations
    - The number of trees that contain only a single node divided by the number of trees with two or more nodes
- Mean, Standard Deviation of Thread Comment Lengths
- Total Number of Trees
- The Depth of the Deepest Tree
- Mean, Standard Deviation of Tree Depths
- Average Branching Factor of Trees
    - The branching factor of a comment is the total number of comments responding to a comment. The average branching factor of the set of all trees is the average of all of the branching factors of all non-terminal nodes (nodes without any responses)

I hypothesized that these properties would serve as a way to quantify the degree of user engagement with a post on Reddit. If a post had a high ratio of stub comments to conversations, it might indicate that people were just saying their two cents worth and moving on to more captivating content. If users were engaging in more lengthy conversations and debate about a particular post (thus increasing the average branching factor, average depth, and maximal depth properties of the comment thread and possibly a lowering the ratio of stub comments to conversations) this could be indicative of the compelling content of the post and a greater possibility of garnering up votes and attention.

III. Fuzzy Cognitive Maps

A. Overview

Bart Kosko developed fuzzy cognitive maps in the mid 1980's. He describes them as, "fuzzy-graph structures for representing causal reasoning." In general they are represented as a digraph. Each node of the digraph represents some causal concept. The value assigned to a node is a fuzzy value indicating the degree of activation of that concept. Each directed edge between each node/concept represents a causal connection between concepts. The weight assigned to the edge is a fuzzy value indicating the degree of the causal connection. FCM's have been used to model and simulate complex nonlinear systems in many application domains including the social sciences, and control systems.

B. Simulating a Complex System Using an FCM

All of the important components and variables of the system are represented as a node in the FCM. A transformation function is used to map the components' real value into a fuzzy value for the model in interval [0,1]. The causal connections between the concepts are assigned a direction and weight in the interval [-1, 1] indicating the degree of causality. The system can then by simulated using the FCM by using a calculation rule to compute the next value of each concept. The general form of the calculation rule is:

$$A_j^t = f\left(\mathrm{k}_1^{\mathrm{j}} \sum_{\substack{i=1 \\ i \neq j}}^{n} A_i^{t-1} W_{ij} + \mathrm{k}_2^{\mathrm{j}} A_j^{t-1}\right)$$

The sum of the value of the other concepts at time (t-1) is multiplied by the weight of the causal connection between them and the concept A_j is taken. This sum is then multiplied by a coefficient k^i that indicates how dependent on the interconnected concepts A_j is. This value is then added to the value of A_j at time period t-1 multiplied by a coefficient k^j that indicates how dependent A_j's value at time t is upon its past value. f is a transformation function that maps the value of its input to the interval [0,1].

If the concepts take on real values in the interval [0,1] it is common for f to be a sigmoidal function:

$$\frac{1}{1 + e^{-cx_i}}$$

If the concepts only take on discrete values, either binary values of {0, 1} or trinary values of {-1, 0, 1}, then a function that maps positive values to 1 and negative values to 0 or a function that maps values greater than 0.5 to 1, less than -0.5 to -1 and all others to 0 can be used respectively.

By iterating through this process one can simulate the behavior a system over time. If the concepts take on only discrete values eventually a reoccurring cycle of finite length will be found. If the cycle is of length one, (e.g. the value of the state vector at time t is equal to the value of the state vector at time t-1) it is known as a static state. If the cycle is of length greater than one it is known as a limit cycle.

C. Python Implementation of a General FCM Simulator

In order to run simulations of processes modeled as FCMs I built a general FCM module in python. The module is made up of the following components:

- FCM class
    - This is the workhorse of the module and instances of the FCM class are used to run the FCM simulations. It also contains a simple pattern recognition function that discovers cycles in time series of state vector data.
- Representation Translation Functions
    - These functions can be used to translate a networkx.digraph representation of an FCM into a matrix representation and vice versa.
- Threshold Functions
    - All of the basic threshold functions are defined in the FCM module, including bivalent functions that map to {0, 1} or {-1, 1}, trivalent functions that map to {-1, 0, 1} and logistic signal functions (sigmoidal) that map to [0, 1] and [-1, 1].
- Random FCM Generator
    - This function generates a random FCM given input parameters that dictate the number of nodes, the degree of connectivity of the FCM, and a function to generate random weight values. The purpose of this function is to generate FCM's for use testing and validating learning algorithms designed to learn FCM's from a set of state vectors.

IV. Modeling Reddit as a FCM:

A. Model Construction

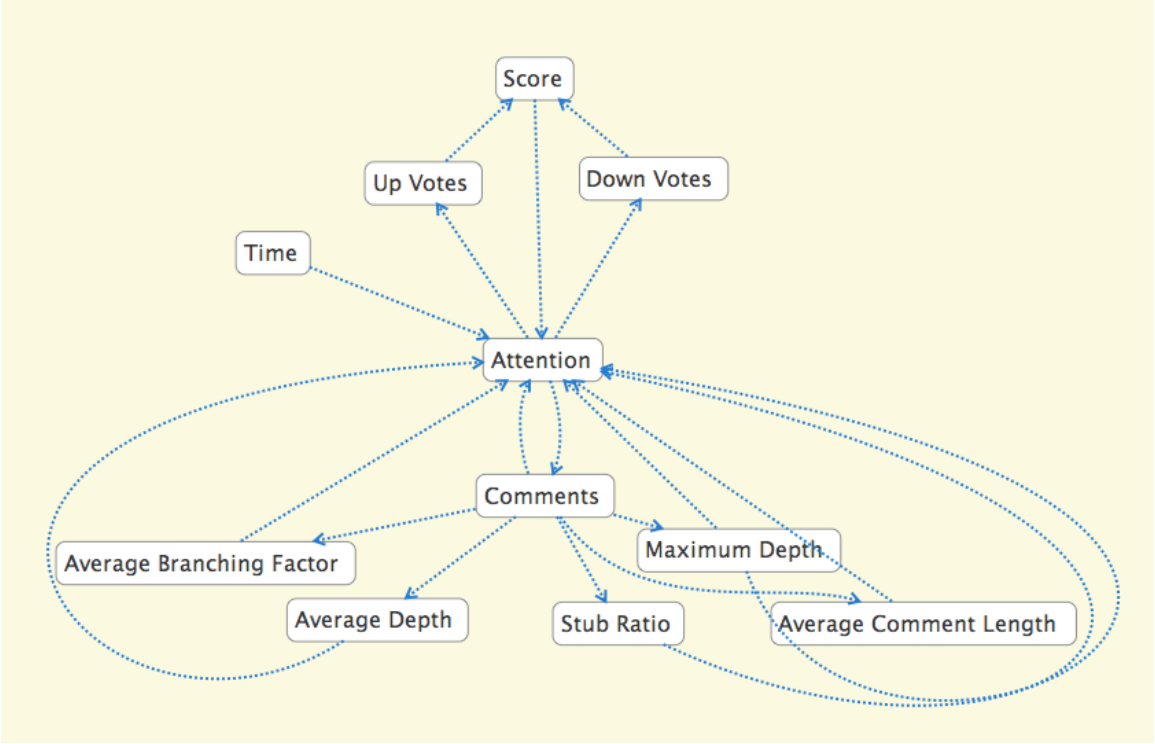
Much of the literature involving FCMs discusses how to construct them using an aggregation a number of experts' descriptions of a system's concepts, edges, and edge weights. Given the complexity of the FCM and my dearth of experts to interview about the possible degrees of causal relationships that exist between the different components of the Reddit system, I used a simple statistical measure of

correlation, the Pearson-R value, to try to determine appropriate weights for the edges in the FCM.

My method of determination was to first split my database of raw Reddit information into two separate databases each containing the data for roughly 800 Reddit posts/comment threads: one for use in constructing my FCM model of Reddit and the other to use to test and validate the model against data not used in its construction. Using the model training database, I aggregated all of the time series data for all of the posts/comment threads and associated properties stored in the CouchDB database. For each data point I then performed a Pearson-R analysis for each pair of properties generating the following table of Pearson-R values

	score	ups	downs	average_branching_factor	average_max_depth	max_depth	total_nodes	ratio_of_stubs	average_comment_length
score	1	0.95	0.81	0.09	0.12	0.28	0.48	0.04	0.04
ups	0.95	1	0.95	0.08	0.10	0.28	0.48	0.04	0.03
downs	0.81	0.95	1	0.06	0.09	0.22	0.44	0.04	0.02
average_branching_factor	0.09	0.08	0.06	1	0.75	0.68	0.31	-0.19	0.25
average_max_depth	0.12	0.10	0.09	0.75	1	0.88	0.41	0.39	0.47
max_depth	0.28	0.28	0.22	0.68	0.88	1	0.68	0.22	0.42
total_nodes	0.48	0.48	0.44	0.31	0.41	0.68	1	0.09	0.23
ratio_of_stubs	0.04	0.04	0.04	-0.19	0.39	0.22	0.09	1	0.38
average_comment_length	0.04	0.03	0.02	0.25	0.47	0.42	0.23	0.38	1

I then used these values as the weights of the edges between concepts and used my judgment as a user of Reddit to determine the direction of the causal relationship producing this FCM:



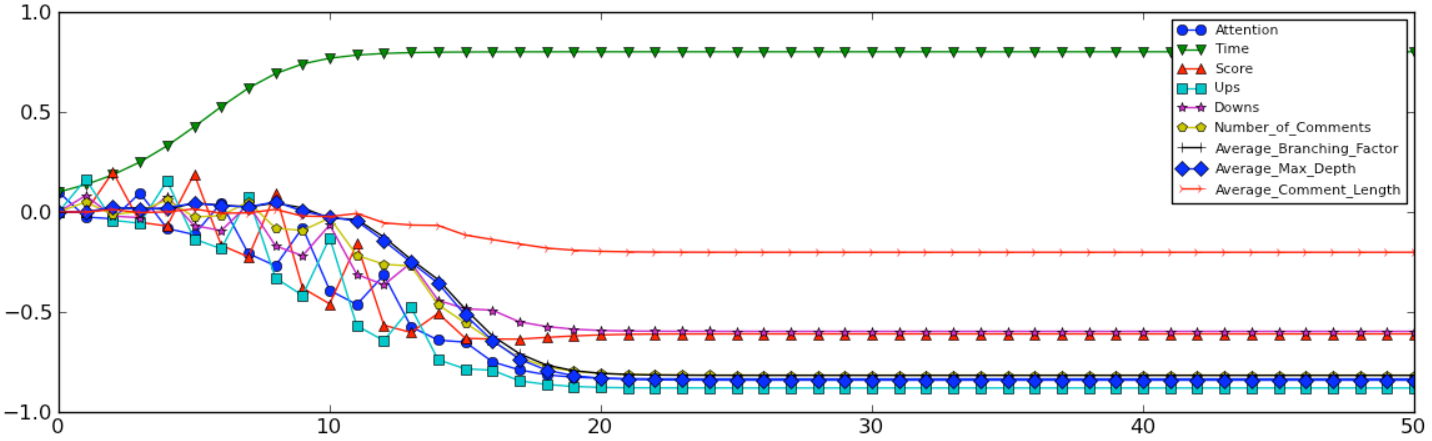
The only concept for which I do not have quantifiable data is "Attention." It is meant to be a representation of the number of times a post on Reddit is seen or glanced at by someone browsing the site. The relative placement of posts on Reddit varies directly with the logarithm of their direct score and inversely with the number of seconds since it was originally posted. The better a post's relative placement the more likely it is to get attention. To add this facet of the Reddit system into the FCM I used a time concept that gradually increases itself (has a positively weighted directed arc to itself) at each iteration and has a negative causal relationship with attention.

I ran 546 simulations using this FCM with varying initial state vectors and weight assignments. There were two main foci of these experiments. One was to observe the effects that the graph theoretic properties described above would have on the overall behavior of the system when they are assigned a causal connection of varying weights to the Attention. The other was to find tipping points in the form of initial concept value assignments. These tipping points are marked by drastically different system behavior given only a slight difference in concept input value. An

example of a tipping point is given below. The graphs show output from the same FCM given two initial state vectors that are identical except for a 0.05 difference in the value of the Attention concept.

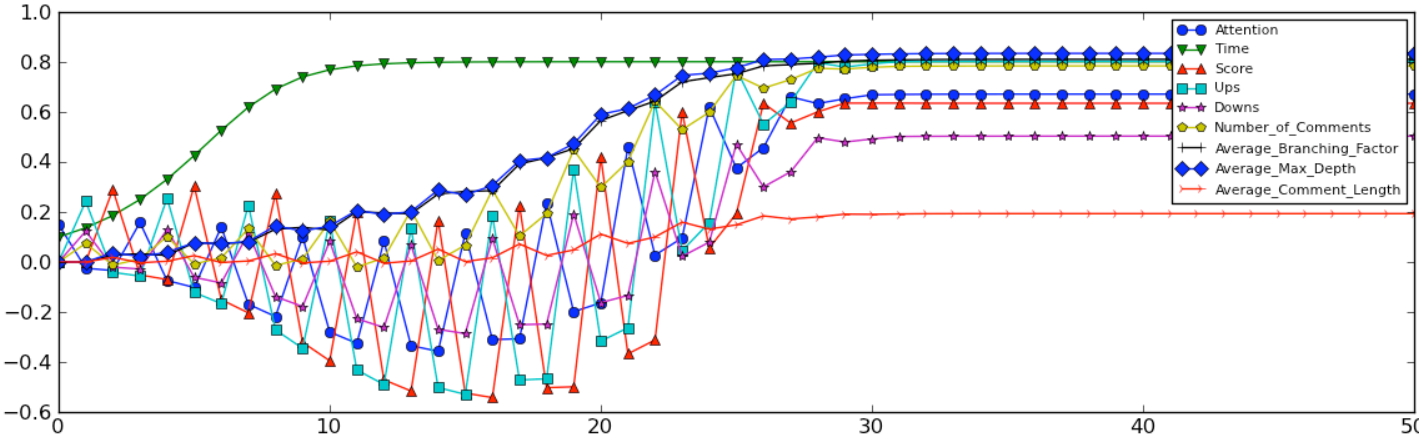
Initial State:

Average_Comment_Length	0
Average_Branching_Factor	0
Number_of_Comments	0.0
Downs	0
Attention	0.1
Average_Max_Depth	0
Score	0.0
Time	0.1
Ups	0

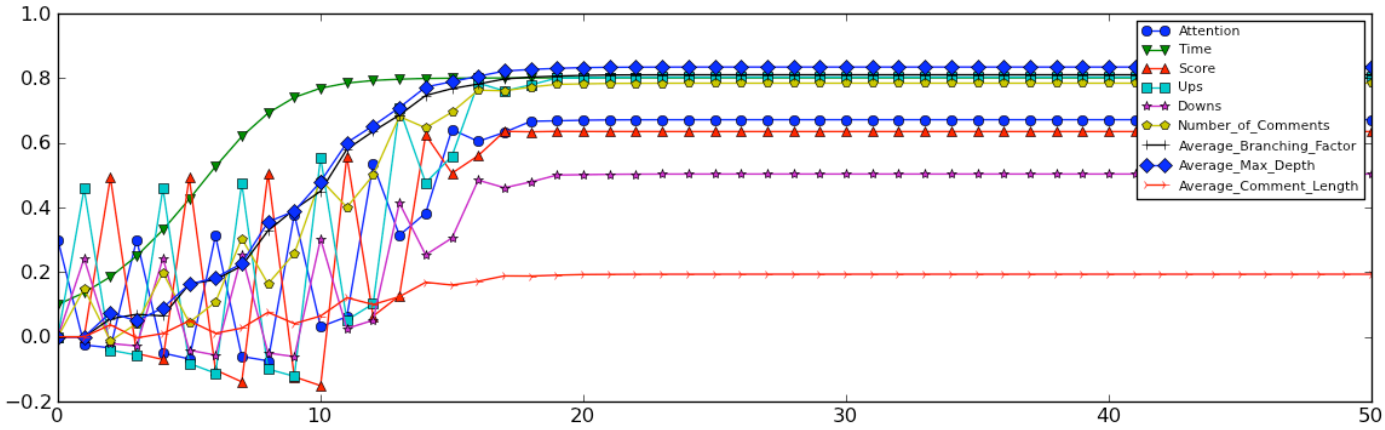


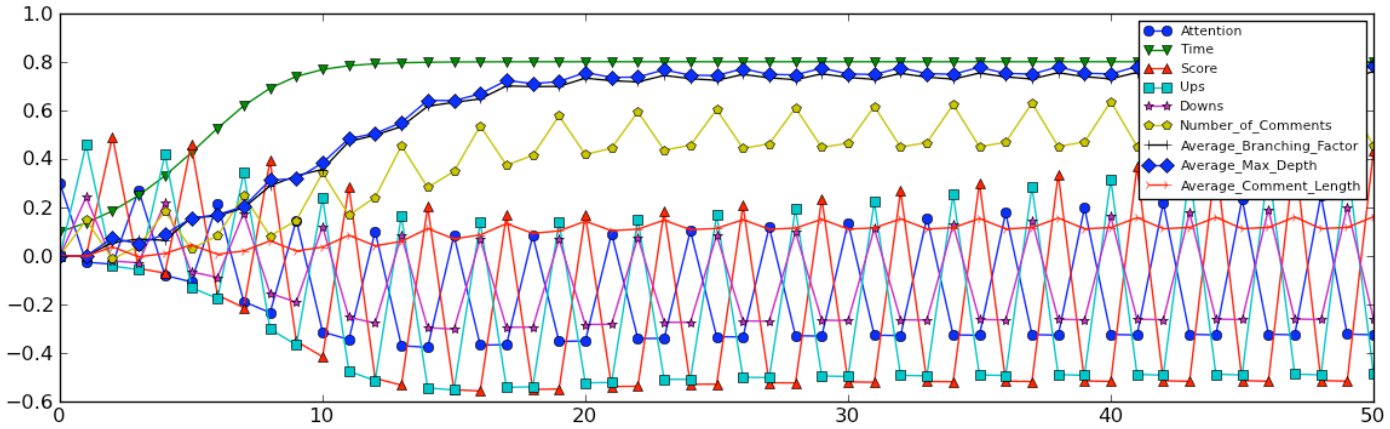
Identical initial state vector except for:

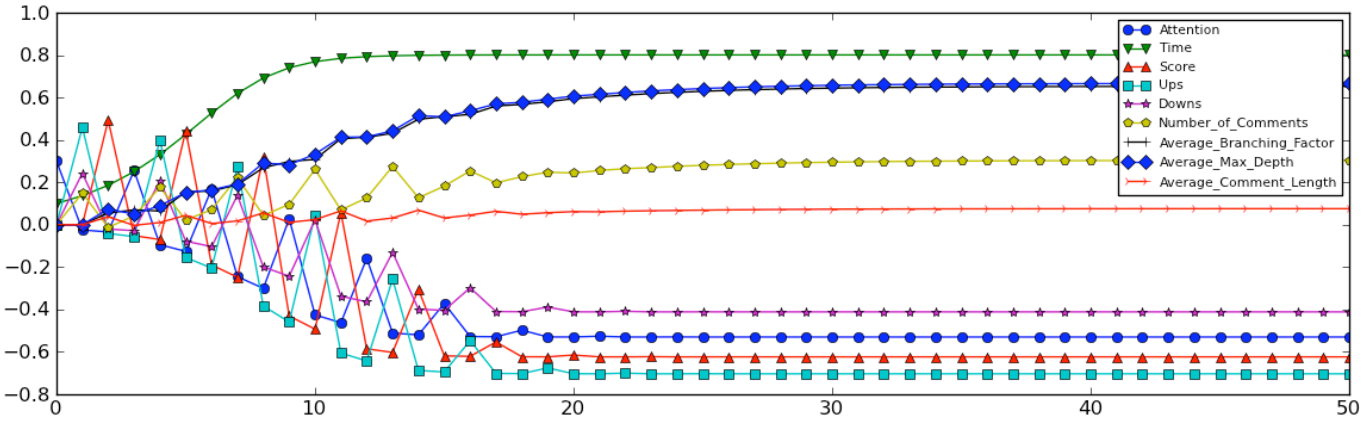
Attention	0.15



Varying the degree to which my hypothesis --that a greater average branching factor and average tree depth positively affected the level of attention a post received-- was true also seemed to generate some interesting behavior. The following graphs show the result of assigning the causal weight between the average branching factor and Attention and the causal weight between the average tree depth and Attention to .15, .05 and 0 respectively while holding all other factors equal.







V. Using a Genetic Algorithm to Learn the Structure of FCMs

Rather than relying on expert knowledge and the manual construction of an FCM it is possible to use machine-learning techniques to generate the causal connections and weights of an FCM based upon historical data of a system. Although experimenting with my hand constructed FCM of Reddit allowed me to generate some interesting behavior I am doubtful as to how effective this approach can be towards generating a truly predictive model of Reddit user dynamics. To this end I implemented a general FCM learner that uses genetic algorithms to discover the causal connections and weights between concepts of a system given historical data generated by that system. This work was closely based upon the research presented in the paper, "Genetic Learning of Fuzzy Cognitive Maps," by Stach et al. in which they use a real valued GA to evolve FCM models to fit known data.
My python implementation of the paper still needs to be thoroughly tested and has not yet been used on the Reddit datasets.

The module containing the FCM Learner is called FCM_GA and contains the LearnFCM class. LearnFCM is a subclass of the genetic algorithm engine of the python library pyevolve and makes use of the general FCM simulator I programmed to simulate the behavior of candidate solutions in the GA's evaluation function. Future work will focus on validating the LearnFCM class on data generated from randomly generated FCM's and then applying it towards discovering FCM's that demonstrate similar behavior to that of Reddit.

IV. Bibiliography

K. Lerman, A. Galstyan, "Analysis of Voting Patterns on Digg," *Proceedings of the first Workshop on Online Social Networks* (2008)

E. Khabiri, C. Hsu, and J. Caverlee,"Analyzing and Predicting Community Preference of Socially Generated Metadata: A Case Study on Comments in the Digg Community," *Proceedings of the Third International ICWSM Conference* (2009)

C.E. Shannon, "Prediction and Entropy of Printed English," *Bell System Technical Journal*, January 1951

Bart Kosko, "Fuzzy Cognitive Maps," *International Journal Machine Man Studies* 65-75 (1986)

K Lerman, R. Ghosh, "Information Contagion: n Empirical Study of the Spread of News on Digg and Twitter Social Networks," *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*

K. Lerman "Social Networks and Social Information Filtering on Digg," *ICWSM'* (2007)

R. Taber, "Knowledge Processing with Fuzzy Cognitive Maps," *Expert S.vstems With Applications.* Vol. 2. pp. 83-87, (1991)

C. D. Stylios, V. C. Georgopoulos, P. P. Groumpos, "Modeling Complex Systems Using Fuzzy Cognitive Maps," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions Issue: 1, 155-162, January 2004

J. Bollen, Alberto Pepe Huina Mao, "Modeling Public Mood and Emotion: Twitter Sentiment and Socioeconomic Phenomena," arXiv:0911.1583v1 November 9, 2009

S.T. Mohr, "Software Design For a Fuzzy Cognitive Map Modeling Tool" 66.698 Master's Project Fall 1997 Rensselaer Polytechnic Institute

D. J. Watts S. H. Strogatz, "Collective dynamics of 'small-world' networks" *NATURE VOL 393 4,* June 1998

R. Crane D. Sornette, "Viral, Quality, and Junk Videos on YouTube: Separating Content From Noise in an Information-Rich Environment" *Association for the Advancement of Artificial Intelligence* (2008)

C. D. Stylios, V. C. Georgopoulos, P. P. Groumpos, "The Use of Fuzzy Cognitive Maps in Modeling Systems" Proceeding of 5th IEEE Mediterranean Conference on Control and Systems, Paphos