Power of Web 2.0 Mass Collaboration in Computational Intelligence and its' uses, an example from Finance

Martin D. Sykora
Department of Computer Science
University of Loughborough
Loughborough, LE11 3TU
UK
M.D.Sykora@lboro.ac.uk

Abstract

One of the main issues of concern within world wide web is the emergence of web 2.0 mass collaboration systems and our understanding of this new phenomenon. Web 2.0 systems have gained enormous popularity in recent years, however as is often case with novel technologies, the real merits sometime stay somewhat obscured to many researchers. In this short paper web 2.0 applications are lightly introduced, with parallels to computational intelligence being drawn and some experimental results from the financial markets presented, to illustrate value of web 2.0. This paper highlights a number of important issues that deserve academic attention. We hope this paper will serve as a light and not too technical introduction towards encouraging others in computational intelligence to consider leveraging web 2.0.

1 Introduction

Everyone has probably heared of wikipedia, facebook, answers.yahoo[1] and terms such as Open source, Blogs or Wikis. These applications have taken the internet community by storm [20, 12, 17], but what do they really stand for, and how, if at all, might they be of value to the field of computational intelligence. These applications are often referred to under the somewhat vague, but yet eloquent term, web "2.0". The term really stands for a major ground shift that started in the last few years and is still taking place. It refers to a perceived second generation of world wide web that facilitates communication, information sharing, interoperability and collaboration on the web. It stands for the paradigm that pretty much everything on the internet is read and write [23], or in some way shareable and easy to distribute. In other words, instead of an Administrator updating a website, now everybody can update, upload and collaborate on content of webpages [20], whether these are file, personal detail or news sharing applications.

As with many new developments, new concepts and twists on using old technology in new ways emerge rapidly and it generally takes time for research to adapt. In this paper we provide some background to web 2.0, related concepts and we introduce an example application of a web 2.0 system to show its potential. The rest of this paper is organised as follows: section 2 looks at mass collaboration, and introduces parallels with optimisation in computational intelligence, in section 3 we provide an example application of web 2.0 from finance and we conclude the paper with section 4.

2 Background

Thanks to implementation of web 2.0 applications, mass collaboration has emerged as it never has before. This occured mainly due to the manner in which web 2.0 applications operate. They are scalable and allow individual users to easily contribute to a problem domain or knowledge base. Generally speaking, we can relate mass colaboration to problems in computational intelligence with the only difference that here the human intelligence is predominantly involved at solving the optimisation task. Whether the fitness function is to minimise forecast errors, or built optimal taxonomies and classes or collect accurate knowledge bases, it does not matter. The next sub-section introduces mass collaboration in greater detail, however, as most readers will realise computational intelligence has for a while been playing around with the idea of incorporating the human element into algorithms.

1. http://www.wikipedia.com, http://www.facebook.com, http://answers.yahoo.com

There are numerous reasons for such an effort, one of the more obvious ones is simply that humans are better at judging subjective and hard to quantify datasets. For example, some GA based art generation work [14, 16, 22, 9, 4], used human subjects to evaluate fitness of machine generated images and music in order to assess the fitness and hence the survival likelihood of an image / musical-piece within a GA, i.e. $Fitness = f(\mathbf{v})$ where fitness is a function of $\mathbf{v}$, which is some vector of "subjective" function parameters. The paper by Takagi [19] provides a very nice and wide literature review of interactive evolutionary computation. Web 2.0 approach to faciliating mass collaboration brings opportunities in at least two forms. First, building custom interactive evolutionary algorithms has become feasable with web 2.0 standards[2]. Parts of or the entire evolutionary process can be outsourced to humans. Secondly, a number of successful applications such as youtube, delicious, or ZiiTrend and Trendio have been built and successfully used to optimise datasets towards some user-perceived fitness. Examples of these are given in section 2.3 and results from our recent experiments that strongly illustrate the power of one of such "web of participation" tools, namely the video sharing platform youtube, is presented in section 3. Our experiments investigated the degree of efficiency in the propagation of financial news within youtube, with highly encouraging results.

2.1 The Mass Collaboration effect

Mass collaboration or what we like to refer to as knowledge optimisation based on intelligence of crowds, and its benefits, whether in terms of problem solving or simply information gathering and filtering, are very appealing reasons behind the adoption of Web 2.0 systems. In a read and write web the contribution of individual users on a large scale amounts to optimisation in the sense of improving a fitness function [20]. Where such a fitness function is understood to be a goal or purpose of the social sharing application, this tends to happen due to emergence of statistical regularities in the evolution of collective choice from individual behavior. To illustrate this idea further, let us look as an example at Wikipedia[3]. Users are encouraged to edit and re-edit this web based encyclopedia in the communal hope of producing an immense body of encyclopedic knowledge. Critics, such as Keen [8] point out the seemingly intrinsic problem, that is such a wast text would clearly have to be riddled with inaccuracies. Quite surprisingly however Wikipedia was found to be an accurate resource and is now becoming a standard encyclopedic reference text. A comparison with encyclopedia Britannica [6] suggests a similar level of information accuracy in both encyclopedias. As was shown, 70%-80% of inaccurate edits on Wikipedia get corrected almost instantly [1, 2]. This can be attributed to the dynamic nature and self-managing environment of collaborative web 2.0. In the case of youtube, numerous users get together to share videos on various topics. Collectively a huge database of videos on wide range of events is built and tagged with meta data. The quality of content is ensured by a mix of expert and non-expert community participants, that review, rate and comment videos. Since it has become so easy to produce and upload videos, youtube has become a feasable application. For further examples and case studies of mass collaboration in practice, please refer to [20].

2.2 Market Predictability

The purpose of example experiments presented in this paper was an investigation into the degree of efficiency in propagation of financial news within youtube. This can be measured by the in-time correlation with stockmarket price data. Price movements in financial markets are consequences of decisions taken by both stockholders and stock buyers based on how they perceive market, sector, company or asset. Actions taken by them are not only influenced by the rational information on market but also what actions other investors took, what somebody said or wrote, and simply sentiment and emotion. According to the recently emerged field of behavioral finance [18], feelings of anger, fear, uncertainty or confidence and subjective perceptions of financial perspectives of economic agents have real impact on entire markets and therefore price movements. In its simplified form, Efficient Market Hypothesis, originally proposed in the 60s [5], essentially states that market participants have equal access to information, and as new information affecting a market comes out, this information is counted-in into the market almost instantly. An offshoot of this hypothesis is the Adaptive Market Hypothesis [13].

2. Standards in this context refer to the technological and conceptual approaches used in web 2.0 systems, see [23]

3. Wikipedia is an online, publicly maintained encyclopedia. It covers millions of topic definitions.

AMH takes behavioral finance into account, and it is within this framework that it is acceptable to expect some short to medium term predictability, based on the information we extract from youtube. This is possible, however, only if assuming information propagates into youtube quickly enough, and can be filtered well from non relevant information. Indeed, we discovered a strong relationship in number of cases. Our findings point to the hypothesis that news data in fact must propagate through youtube quite efficiently, see 3 and 4.

2.3 Examples of Optimisation

In this subsection we present a wider range of examples where popular web 2.0 applications can be considered to have evolved towards certain subjective fitness. The object of evolution is often quality or appeal of the data, however in example of trend prediction platforms the fitness is perceived as minimum squared error between predicted and real values. This list is by far not exhaustive but it never-the-less serves to illustrate some examples.

- Bookmark tagging (e.g. delicious.com) is a popular type of application where any resources on the world wide web can be tagged by users en-masse. These tags are very simple and occur in the form of one-word tags that describe a resource well enough, such as "crisis", "crash" or "recession". Surprisingly enough, this simplicity essentially creates large and optimal taxonomies [15]. Since everybody is involved in the taxonomy building, we often refer to them as "folxonomies" [20].

- Social picture sharing applications such as the well known applications flicker or picassa allow users to submit images, tag, rate or leave comments on them. Since these ratings refer to certain aspects of the visual appeal of images, this picture rating process used with tagging for example provides for subjective measure of image fitness in various folxonomicaly created categories.

- Wikinews, technorati, or reddit are examples of text fitness optimisation platforms. Essentially these applications stand for something that has been in past refered to as democratising news [20], as individual users get to vote for and decide on how high news articles appear in the headlines.

- Trend prediction applications such as ZiiTrend, Trendio or GloboTrends on contraty allow users to vote on what they believe will be future outcomes of events. The two interesting aspects to this is that accuracy of predictions can be measured accurately (hence users with good prediction skills can be identified for example), and that objects of predictions are suggested by users themselves (users select topics of interest that are to be predicted).

- Optimisation of social networks also occurs on platforms such as orkut and facebook. These networks have been shown to have some interesting properties and a number of researchers investigated these, some of them [10, 11, 7, 3]

In the next section we introduce in more detail the results of an investigation into efficient news information transfer within the media sharing platform youtube. The case of media sharing is a more special situation, since submission of video clips is a more elaborate process than simply tagging or voting for data that is already online. It is hence interesting to see whether the concepts of optimisation towards quality / usefull content still works, even if the process requires more effort on the user side. In order to get a realistic idea of how usefull the content really is, principles of financial markets were used.

3 Experiments and Results

The goal of our experiments is to show there is correlation between changes in stockmarket prices and community submitted information on the youtube platform. We chose this web application for a number of reasons. First of all it illustrates the power of mass collaboration to a fitness and our results demonstrate successfull application of making use of it. We could model the fitness function of a video sharing community in detail, however for the sake of clarity and brevity let us assume; fitness depends on the usefullness of a video and the usefullness of it in the "financial news" further depends on the video containing the most up-to-date and highest quality video content. Our hypothesis is, since most popular news videos receive attention, information gets propagated through youtube quick enough to satisfy efficiency expectations (see section 2.2) but more importantly this information represents value in terms of capturing financial news events that can

be correlated to the markets. In order to show this relationship we took a set of steps to obtain, extract, prepare and analyse youtube data. We were interested in answering two questions. First, whether it is possible to relate intensity of content submissions with market volatility. Secondly, whether it is possible to quantify sentiments of videos and relate them to directional market moves.

Youtube.com was established in February 2005, as an online video and the premier destination (..) to watch and share original videos worldwide through a Web experience[4]. It is a free community-driven website through which registered users can upload unlimited number of videos and share them with other users. Each video must be given a title and be assigned to a specified category (e.g. News, Music). A publisher can optionally provide further details. According to alexa.com, web traffic has been constantly growing since its founding, earning youtube a ranking in top 3 most frequently visited websites in the world. It reaches about 5% of internet users in a day and generates 20% of all http based pageviews on Internet. These figures make youtube most popular community based website.

3.1 Input Data

Every uploaded video on youtube is in the form of a video file and a set of related meta data describing the file. Such meta data contains video title, description, category, date of submission, view count, duration and author. Since youtube is a social website it also allows users to comment, rate (1 out of 5) and submit response videos. Videos can also be tagged with arbitrary tags that might help identify a video better. Viewcount and Comments are quite important for our analysis, the former can be important in judging the popularity of a video and the latter also provides us with collective opinion about a videos' contribution in textual form. Title, Description and Comments are all textual information streams and once these are serialised in time-stamp order, they can provide information about evolution of market sentiment. Financial news videos based on search keywords FTSE, DOW JONES, NASDAQ, NIKKEI, CAC and DAX[5] were retrieved. The analysis was performed on around 90 thousand videos and over 3.5 million comments. Textual data was preprocessed for text analysis, (i.e. tokenising, stemming, stop word removal, emoticon handling, etc.) and financial indices were smoothed by a windowing based time-series segmentation method, in order to remove noise trading and other disturbances in the underlying price data.

3.2 Sentiment Analysis

Initial investigation showed that frequency of video submissions is strongly correlated with stock prices for the US markets. The Pearson correlation coefficient for the whole period January 2007 - January 2009, was 0.745 for video submission and 0.697 for comments submissions. Both values are statistically significant and point towards a relationship between stockmarket and youtube.

Further three vocabulary based models were built to quantify text sentiment. These models output a score s based on scoring function of the form:

$$s = \frac{p - n}{p + n}$$

where p is number of positive words and n is number of negative words in the text, $p \geq 0, n \geq 0, -1 \leq s \leq 1$. This scoring function is relatively intuitive and self explanatory. Even thought quite simple, it is robust in capturing word bias in a piece of text, and has been used by other researchers in the past [21].

When aligned against stock index returns, correlations of 0.423, 0.387, 0.033 were measured for title, description and comment models respectively, where the first two correlations are statistically significant. These varying strengths of correlations are due to the fact that there is noticeable difference between the three streams of text. There are a number of reasons for this. Comments tend to be relatively noisy, and filtering them from noisy contributions can be difficult, hence their use in judging sentiment is somewhat obscured. Improved results were achieved when the scores were combined into a single indicator by averaging the individual scores. See Figure 1 for this statistically significant 0.543 correlation. As one can see, the resulting score tends to correlate in local turning points to the market very consistently.

Finally we employed thresholds to the combined score, in order to change soft classification to proper classification. Since scores are distributed normally with a mean (μ) of -0.021 and variance (σ^2) of 0.003721, we used this to eliminate some of the more frequent values close

4. According to youtube http://www.youtube.com/t/about?hl=en_GB (consulted on 2 April 2009)

5. All financial indices.

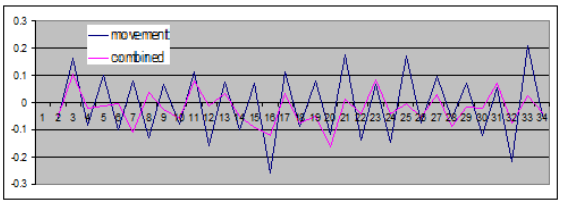

Figure 1: Combined (mixed) scores and segmented price data. *(x-axis monthly time-line, y-axis normalised combined score and normalised monthly stock returns)*

Table 1: Classification of how often the sign of score is in agreement with directional market movement

Scenario	Lower and upper limit	Matches / Hits	Hit rate
No limits	None	25/33	76%
$\mu \pm 1/4\sigma$	-0.0367 to -0.0059	21/25	84%
$\mu \pm 1/2\sigma$	-0.0521 to 0.0094	16/18	89%
$\mu \pm \sigma$	-0.0829 to 0.0402	6/7	86%

to the mean. Table 1 illustrates the rather good (76% up to 89%) model accuracies of directional move forecasts.

4 Conclusion

In this paper we introduced web 2.0 in context of computational intelligence and touched upon its possible uses within the field. A world wide web based on web 2.0 allows to leverage human subjective judgmental abilities. In many real world situations it is difficult to build models that mimic human decissions. Since it has become feasable to tap into human involvment on a large scale, the human factor can now be integrated into computational algorithms with relative ease. In this paper, the immense power of mass collaboration was highlighted and introduced by an example of our recent work on stock market news efficiency as distributed over a web 2.0 media sharing platform of news content sharing on a large scale. It was shown that strong correlation with (mass user submitted) financial news media and the main markets is statistically significant. We further illustrated a predictability of market direction based on sentiment with accuracies ranging between 76% and 89%. Given current theories of financial market efficiencies (section 2.2) it could be inferred that youtube as an information sharing platform is rather optimal at propagating news articles that matter. This is thanks to individual users submitting, editing[6] and monitoring[7] what they perceive as valuable content, and in turn "unknowingly" (without even realising it) they collaborate on a large scale to create database of videos that satisfy a perceived quality and relevance fitness. In other words, youtube platform enables the creation of a high quality database of videos, where high quality is defined as a subjective function of quality and overall relevance to a certain topic. This and many other web 2.0 systems are in fact examples of interactive evolutionary computation in action. There is much more to be desired in terms of research within computational intelligence, where web 2.0 platforms are used more directly in implementing IEC[8]. There is also much desired in terms of a unifying framework for such approaches. The current work and results are encouraging and we hope this paper serves as valuable introduction to this relatively new sub-topic.

Acknowledgements

I am very grateful to Marek Panek, who helped to develop and implement a great deal of youtube experiments reported on in this paper. I am also thankful to Dr. Helmut Bez, for his constant and attentive supervision.

References

[1] B. Adler, K. Chatterjee, L. de Alfaro, M. Faella, and V. R. I. Pye. Assigning trust to wikipedia content. In *WikiSym 2008: International Symposium on Wikis*, 2008.

[2] B. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the wikipedia. In *WikiSym 2008 - International Symposium on Wikis*, 2008.

[3] D. Boyd and N. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13:210–230, 2008.

[4] S.-B. Cho. Towards creative evolutionary systems with interactive genetic algorithm. *Journal of Applied Intelligence*, 16(2):129–138, 2002.

6. Clips can be rated, commented on and replied to by other videos.
7. Spam clips can be reported by users.
8. Interactive Evolutionary Computation

[5] E. Fama. Random walks in stock market prices. *Financial Analysts Journal*, 51, 1965.

[6] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.

[7] S. Golder, D. Wilkinson, and B. Huberman. Rhythms of social interaction: Messaging within a massive online network. In *Communities and Technologies 2007: Proceedings of the Third Communities and Technologies Conference*, 2007.

[8] A. Keen. *The Cult of the Amateur: How the Democratization of the Digital World is Assaulting Our Economy, Our Culture, and Our Values.* Doubleday Currency, 2007.

[9] A. Kosorukoff. Human based genetic algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, 5:3464–3469, 2001.

[10] C. Lampe, N. Ellison, and C. Steinfield. A face (book) in the crowd: Social searching vs. social browsing. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 2006.

[11] C. Lampe, N. Ellison, and C. Steinfield. A familiar face (book): profile elements as signals in an online social network. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 435–444, New York, USA, 2007.

[12] C. Li and J. Bernoff. *Groundswell: Winning in a World Transformed by Social Technologies.* Harvard Business School Press, 2008.

[13] A. Lo. The adaptive market hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management*, 30:15–29, 2004.

[14] M. Ohsaki, H. Takagi, and K. Ohya. An input method using discrete fitness values for interactive ga. *Journal of Intelligent and Fuzzy Systems*, 6:131–145, 1998.

[15] L. M. Orchard. *Hacking Delicious.* Wiley, 2006.

[16] G. Papadopoulos and G. Wiggins. A genetic algorithm for the generation of jazz melodies. In *Proceedings of STeP*, 1998.

[17] C. Shirky. *Here Comes Everybody: The Power of Organizing Without Organizations.* Penguin Press, reprint edition edition, 2009.

[18] J. J. Siegel. *Stock for the long run - Guide to Financial Market Returns and Long-Term Investment Strategies.* McGraw-Hill, 3rd edition, 2002. ISBN: 0-07-137048-X.

[19] H. Takagi. Interactive evolutionary computation: fusion of the capabilities of ec optimization and human evaluation. *Proceedings of the IEEE*, 89(9):1275 – 1296, 2001.

[20] D. Tapscott and A. D. Williams. *Wikinomics - How mass collaboration changes everything.* Atlantic Books, 2008.

[21] P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62:1139–1168, 2007.

[22] N. Tokui and H. Iba. Music composition with interactive evolutionary computation. In *Proceedings of the thirrd International Conference on Generative Art*, 2000.

[23] G. Vossen and S. Hagemann. *Unleashing Web 2.0: From Concepts to Creativity.* Morgan Kaufmann, 2007.