# Analyzing Social Media Tweets with SAS®

Sy Truong, Meta-Xceed, Inc. (MXI), Fremont, CA


## ABSTRACT

The library of congress is archiving all Twitter messages since the messages posted capture the most unfiltered and immediate opinions representing the ethos of popular culture. The information posted on Twitter expresses the most up to date pulse on a large range of topics and is a powerful research tool. With several billion messages posted since 2006 along with constant new posts; this creates challenges in finding the information matching your particular interest. This paper describes techniques on how SAS macros is used to analyze the messages or tweets that you posted and compares it with all other posts. It describes how PROC CLUSTER is used along with other SAS logic to find the best correlation of ideas and topics that are of similar interest. It then presents this recommended feed to you. The analysis of SAS goes beyond an individual ability to manually search since the criteria are constantly updated with what you post and what you decide to read from the recommended feeds. Similar to how Amazon.com recommends a book or CD based upon purchase patterns, this paper presents a method on how to analyze your interest through your social media messages. It then recommends others to you based on similarities in topic and content.



## INTRODUCTION

Social media consists of messages from users within virtual communities shared amongst each other. At first glance, the messages appear to be insignificant since they are short, filled with commentary and links to other messages or websites. To the uninitiated, this seems to be trivial teenage chatter including self absorbed individual declarations about what they had for lunch. Old media requires a voice that has gravitas and authority such as the New York Times or CNN. This media landscape was caught by surprise by the rise in popularity of new media which consist of personal, individualized communication that has old media scrambling trying to recapture their audience. What old media initially perceived to be insignificant noise has crescendo into a force that displace the economics of advertising as marketing of products and services shifts focus to user reviews rather than the traditional broad catch all commercials or newspaper advertisements. Besides advertising, social media can also deliver breaking news as messages are sent directly from cell phones rather than broadcast network television or newspapers. Social Media is encroaching upon a wide spectrum of content ranging from entertainment, news to advertisement. Traditional media is being augmented if not replaced by social media as this new media is capturing more people's attention overshadowing even the use of search and email on the Internet.

Social media relies on a couple of constructs that has only recently come into being within the Internet age. They appear to be on opposite ends of the spectrum including the "Wisdom of Crowds" and the "Long Tail". Both of these forces complement each other in validating and ensuring the value of the content that social media is churning out. James Surowiecki wrote the book entitled "The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business". His book argues how the opinions of many as it is aggregated can lead to a higher accurate representation of the truth as compared to an individual opinion. An example of this is if a crowd of people are at a fair and there was a jar of jelly beans that was fairly large filled to the top. People of all ages at the fair would take a guess as to how many beans are in the jar. Some guess too low while others overshot their estimates. However, if the average of all the guesses were tabulated, it would be the most accurate proving his theory of wisdom of crowds. His examples are primarily illustrated within economics and psychology but it can also be applied to the social medium capturing the most accurate snapshot view for our society. Some examples include the dissemination of news or product reviews.

Another phenomenon that has existed for a while but became popularized when Chris Anderson wrote an article on Wired magazine describing how the Internet has allowed merchants such as Amazon.com and Netflix to provide many niche products that only a few people would buy. Since the Internet could provide a virtual warehouse of products that is more efficient and larger than traditional brick and mortar stores; little unknown books, movies and music can be sold more efficiently. Anderson argues that the many specialized products if graphed would show a

very long tail. The long tail is so long that when added up, it becomes more profitable and larger than the "head" of the distribution curve. He describes that the days of the blockbuster artist like Michael Jackson or Modanna can no longer be as profitable today as compared to the 80s since consumers can now easily find esoteric artists that before would not have any audience. This principle can also apply to social media in the same way that consumers are no longer limited to three television networks or a couple of news papers but hundreds of thousands of blogs, Facebook and Twitter messages discussing a wide varied set of topics constantly.

Social media is also different in that it is interactive allowing individuals to produce and consume information simultaneously in a way that was not available before. The deluge of information has prompted smart social media users to aggregate and get to the information that contains the most meaning to them as described in the Wisdom of Crowds. The abundant of different information sources that was not available before has transformed information into many distinct customized feeds for many individualized tastes and niches. This re-distribution of information among many distinct social networks is what leads to the long tail affect. The Long Tail has a democratizing affect on information as it allows individuals to rival and replace traditional reporters within the traditional media outlets. The combination of the Wisdom of Crowds and the Long Tail which I will refer to as the "Wise Tail" is a perfect storm combining and enabling an efficient dissemination and consumption of information culminating in a new form of media that is more valuable to consumers compared to old media. Users can now arm themselves with tools to navigate the information finding the "channels" that contains the most pertinent and meaningful to them while also allowing them to interactively express their opinion and ideas to a community that is receptive. This paper describes how SAS can empower the "Wise Tail" by analyzing social media content and making "Wisdom of Crowds" suggestions on all topics within the Long Tail of information on the Internet.

## SAS MACROS FOR SOCIAL MEDIA ANALYSIS

SAS has many powerful analytical tools along with a large array of methods to utilize to analyze social media content. SAS recently released a new software product referred to as "SAS Social Media Analytics" which does a good job at analyzing the semantics of social media content to determine if people are saying positive or negative things about a particular product or brand. It uses social media content in a very different way than what this paper will be demonstrating. This paper will use SAS to aggregate social media content for the purpose of individual consumption rather than having it monitor a brand or advertisement campaign for a business. The examples illustrated uses SAS macros rather than applying SAS Social Media Analytics solution.

There are many user friendly tools to help analyze text such as using JMP or EG (Enterprise Guide) to assist in developing the logic for social analysis. Upon review, I found that the most flexible and direct way to get to the full power of the SAS system for a very specific task such as analyzing social media content; I used data step, some SCL and encapsulate this with a SAS macro as a form of an API (Application Programmer's Interface) in order to easily utilize these tools repeatedly in flexible way. The macros used in this project include:

1. %twituser - Capture all the Twitter messages from specified user.
2. %twitsearch - Search for the Twitter messages with specific criteria.
3. %twitlike - Recommends Twitter messages that are similar or like the ones from the user specified.

The macro names start with the word "twit" which is referencing the fact that it is accessing Twitter messages. This is one social media outlet but the same tools can be extended to Facebook and other social media sites as well. The macros will be used by logic on the server which can be delivered to a web front end. The development of a social media website will be referenced but is beyond the scope of this paper. The description of how the social media content is captured, analyzed and then aggregated into recommendations is what will be covered from SAS macros. The macros can be applied in three steps.

### STEP 1 – Capture User Messages

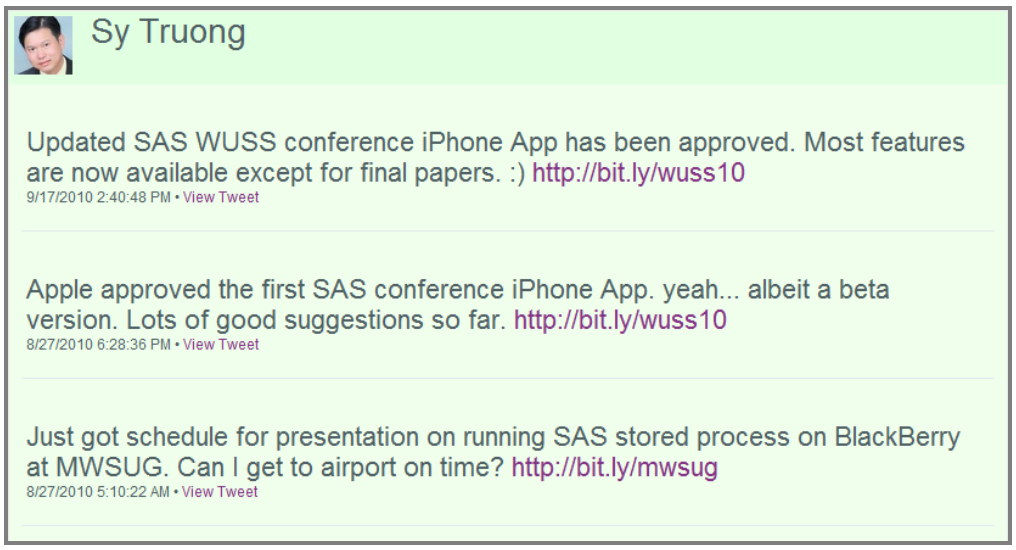
The first step in making sense out of the Twitersphere or world of social media is to understand what messages you have sent out. The messages you have created either through your cell phone or on various Twitter clients such as that from a web browser or iPhone App, are stored on the Twitter website within your own profile. This step would gather your messages and have a collection of the most recent messages you sent out in order to then find others like minded messages presented as recommend ions for further reading. This is accomplished through a macro named %twituser which has the following parameters.

```
%twituser  (user = twitter user name,
            outfile = output report);
```

| Where | Is Type... | And Represents... |
|---|---|---|

| | | |
|---|---|---|
| *user* | C (200 chars) | The user name for the Twitter account. An example is: sytruongus |
| *outfile* | C (200 chars) | The output HTML report listing out all the tweet findings from the specified user. This includes full path with file name. If none is specified, it is defaulted to the current path with file name:<br>twituser.html<br>An example would be:<br>c:\mypath\sytruongus.html |

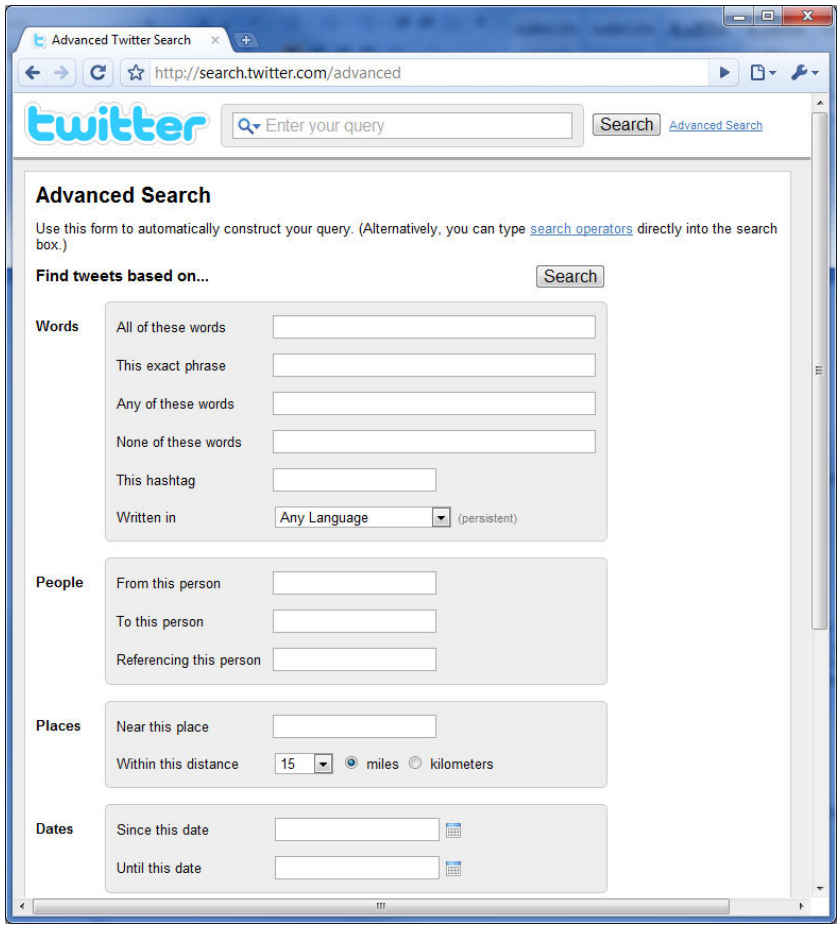
This macro uses the Twitter API and captures the latest Twitter messages for the specified user. The information is stored in a dataset for further analysis and an HTML report is also generated as shown here.




A style sheet is used so the output is displayed in a particular font and size. This can be modified and customized to match other reports or websites within your environment.

**STEP 2 – Search**

Once a collection of all the latest messages from the current user has been captured, the next step is to search what other messages exist on Twitter that are similar to the current user. The Twitter engine provides an API that allows a SAS program or other external tools to perform a search upon specific conditions on Twitter and return all messages that meets the search criteria. You can perform this search interactively if you were to go to the website: http://search.twitter.com. The advance option describes some of parameters you can specify in your search.




The following four options are normally applied during a search.

1. **Key Words –** By analyzing what the currently user most commonly wrote about using PROC FREQ, you can then apply a search to identify other discussion containing these key words.
2. **Language –** In this example, English is used since without this filter, there are many messages that may contain the same keywords but is indecipherable since social media is truly a global phenomenon.
3. **From this Person –** A filter is applied to not include the current user in the search result. This is because the goal is to find messages from other users of similar interest excluding the current user.
4. **Dates –** This can be adjusted depending on how many results are returned. If the topic and search criteria are so esoteric that only a small amount of results are found, the dates can be extended but by default, the search is within the current week.

There are many other conditions that are applied within the SAS program logic to decipher and analyze what is a meaningful message but it is nice to have some of the stipulations applied right up front with the search engine provided from Twitter. This helps narrow down the results so the starting point of analysis is more focused.

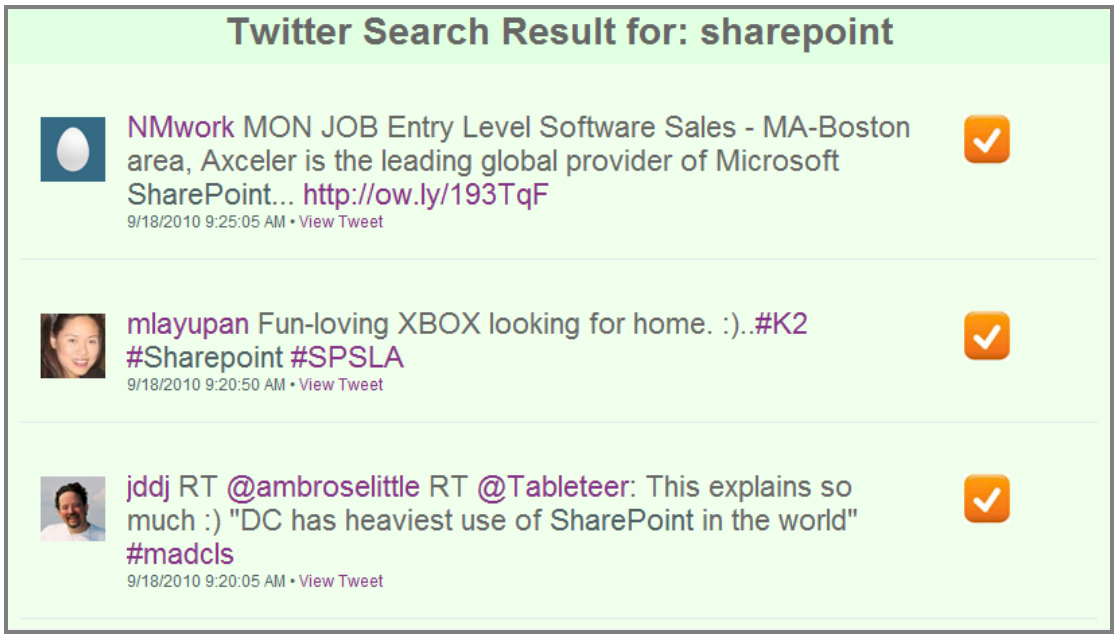
A search is applied through a macro named %TWITSEARH. This is the API for SAS programmers since it can be used as an interface to the Twitter search.

```
%twitsearch (criteria = words search criteria,
             user = from this user,
             date = since this date,
             outfile = output report);
```

| Where | Is Type... | And represents... |
|---|---|---|

| | | |
|---|---|---|
| *criteria* | C (80 chars) | The search text criteria to be applied to http://search.twitter.com. This can be one or more words. This is applied case insensitive. |
| *user* | C (200 chars) Optional | A valid Twitter user name that the search can be applied to. |
| *date* | C (10 chars) Optional | The date in which the search results will be filtered to include. The format is yyyy-mm-dd. An example is: 2010-05-14 |
| *outfile* | C (200 chars) | The output HTML report listing out all the search results. This includes full path with file name. If none is specified, it is defaulted to the current path with file name:<br><br>twitsearch.html<br><br>An example would be:<br><br>c:\mypath\sytruongus.html |

The search results are stored in a SAS dataset. So for example, if I were to search for a text criteria such as "SharePoint", the HTML report returned would look like the following:




The list of findings is shown with the same style sheet used in the %TWITUSER macro so it shares the same font and colors. There is an additional check mark button icon to the right of each search result. This is a feature that allows you click on and select the messages that you like. By identifying which ones you like, the software can make more intelligent suggestions for future messages in which you "like" to read. This is explained more in the next section regarding %TWITLIKE macro.

**STEP 3 – Identify Similar or Like**

Capturing the user messages and then applying searches are straight forward tasks that can be applied by any other programming language or tools other than SAS. Where SAS provides tools that others cannot is the ability to analyze the text and figure out how similar they are based upon various statistical models. In this example, PROC

CLUSTER is used to analyze the text by placing the text into clusters and then finding out their relative distances in order to identify messages that are similar. The macro that is used to identify these like minds type of messages are applied through a macro named %TWITLIKE. This macro contains logic that uses the other two macros in order to find the relevant information. The parameters are therefore simplified to only contain the essential parameters.

```
%twitlike  (user = twitter user name,
            outfile = output report);
```

| Where | Is Type... | And represents... |
|---|---|---|
| *user* | C (200 chars) | The user name for the Twitter account. An example is: sytruongus |
| *outfile* | C (200 chars) | The output HTML report listing out all the tweet findings from the specified user. This includes full path with file name. If none is specified, it is defaulted to the current path with file name:<br><br>twitlike.html<br><br>An example would be:<br><br>c:\mypath\sytruongus.html |

When these three macros are combined, it creates a very powerful and dynamic stream of information to the user customized to their individual interest. This can be executed in batch mode periodically and the resulting HTML reports can be placed on a web server so that the user can view through a URL. Each time the reports are generated from the %TWITLIKE macro, the results are different and this becomes a dynamic report. The factors that make the result change include the following.

1. **Twittersphere Updates –** The resulting report from %TWITLIKE is summarized by searches upon messages on Twitter applied upon a window of time. This window is constantly being shifted with new messages being added by users so the results are constantly being updated creating a very dynamic and fresh view for the user.
2. **User Updates –** If the user creates a new message, this alters the aggregate of keywords that are used to then have the search applied in %TWITSEARCH. Since the search criteria within the keywords change due to the user updates, the results are sure to be updated and different.

Over short periods of time such as a week, the report can be very different. If however, the user is constantly refreshing the report, it may not change radically minute to minute. If the user has the option to select the item they are reading from the feed that they like more than others, this resulting list of "liked" message can also be added to the aggregate of comparisons which can lead to an even more dynamic and accurate correlation between what the user is reading and what the recommendations are providing. This "like" selection is described in the next section as a button within an interactive web site.

The macro use a series of steps identify the list of messages that you "like" to read. It does the following to accomplish this.

**STEP 1 – Identify Current User Key Words**

It first uses %TWITUSER to find all the messages within the current user. The then parses through each word and removing URL and trivial words that are not significant such as "a, the, at, etc….". It uses PROC FREQ to then identify key words that are used the most.

```
proc freq data=wordlist;
   tables word / out=FreqCnt;
run;
data FreqCnt;
```

```
   set FreqCnt;
run;
proc sort data = FreqCnt;
    by descending percent;
run;
```

Once it ranks the keywords from the current user, it starts to search using this for the next step.

**STEP 2 – Search Twitter with Key Words**

From the list of top ranked keywords from the current user, it then methodically search Twitter using %TWITSEARCH for the most recent posts containing these keywords. It stores all these results in SAS datasets to be used for the next step.

**STEP 3 – Identify "Like" Minded Messages**

It compares the results from the search against the current user messages as initially collected in step 1. It starts initially identifying clusters of messages that have similarity using PROC CLUSTER.

```
proc cluster data=test1 outtree=cluster1 method=average standard;
   id name;
run;
```

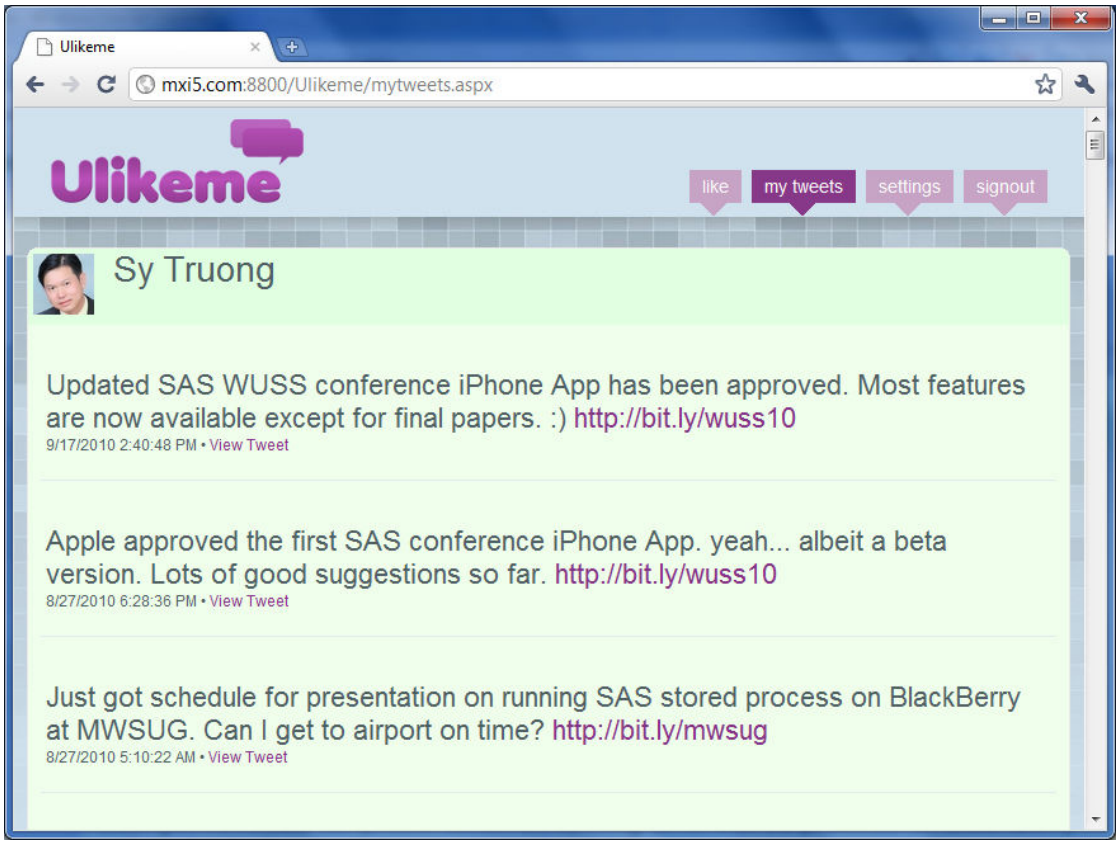
By analyzing the output and identifying clusters that share similarities and are close in distant to each other, it then does a ranking using the COMPGED function. This helps score the messages from the search results which are most closely related to the current user's messages.

```
data likefindings;
   set &datname(rename=(content=content1 twitid=twitid1))
   end=eof1 nobs=nobs1;

   gedscore=compged(content1,content2,&maxscore,'iL' );

   if gedscore > 0 and gedscore < &maxscore and compress(name) = ""
   then output;
run;
```

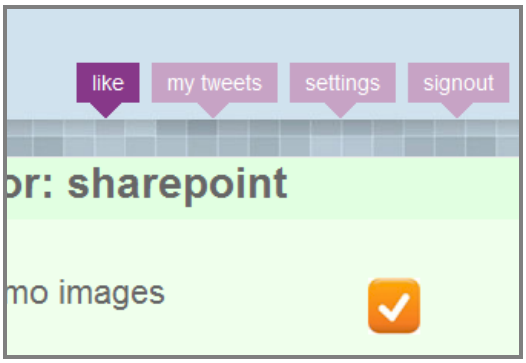
There is a cutoff point based on the `&maxscore`. Once the top related messages are identified to be most similar to the current user, it then presents this in the report which the user can read to further identify which item they like.

## ULIKEME SOCIAL MEDIA WEBSITE

The reports generated from the macros can be invoked on a server dynamically to from a website which is going to be named "Ulikeme". This message displays the same results that are generated from the macros. The differences that it will be wrapped around an interface with navigational links above to allow you to access the results of the various macros.




The “my tweet” screen lists out an expanded view of what you have entered through the Twitter website. This is the basis in which the analysis and searches are applied for the final “like” screen which is the result of %TWITLIKE macro. The one feature which the like screen contains which the macro does process is the “like” button next to each message.




Each time a user reads an article and they find it interesting, they would click on the “like” button. These messages are complied and used in a similar way to that of the “my tweet” screen. The combined findings from correlated messages from the user messages and the ones they liked is then used to derive new recommendations. This acts like a feedback loop that increases the accuracy of the recommendations as it harnesses the “Wise Tail” to get specifically to the discussion and topic of your interest. This makes the experience even more interactive and dynamic because each time you click on a like button, it changes the result.

**CONCLUSION**

Social media has dramatically changed the way users consume information since they are spending more time on social media sites such as Facebook and Twitter as compared to even the most popular search engine site like

Google. Google is a transformative Internet phenomenon that has altered the world of information and is the topic of many books and documentaries so for social media to surpass Google in the amount of time that users spend is a significant cultural shift. The power of social media is that you can trust your group of friends to recommend to you what news to read or what movies to watch. This relies on the concept of "Wisdom of Crowds" in that if you have fifty friends and they all highly recommend a movie, you would be wise to follow this advice compared to a movie critique who may have a very different academic film school perspective. If you have very specific taste and are interested in a topic that only a small handful of people partake in, you can still find friends on social networks since the user base are vast creating a very "Long Tail" in the distribution curve on any topic. Social media can combine these elements in forming the "Wise Tail" which leads to even better and more accurate information. Currently social media sites can figure out the linkage of friends and recommend other friends to you enlarging your social network. The idea is that if you are good friends with John and John is good friends with Joe, then Joe would likely be compatible friends with you. This may work within small circles but as this expands and as your interest diverges into niches; your social network may not be optimal. This commonly happens because as human beings, we have different social networks such as one as a parent at home and another as a professional at the office. The professional versus personal is one example but we normally inhabit tens of hundreds of distinct networks due to our many interests. The goal the macros and software described in this paper illustrates how you can fully harness the power of a global social network expanding beyond the constraints of the interactions we have in the physical and the restrictive "friends" construct that exists on most social network sites. This paper fully utilizes the "wise tail" by expanding and exposing you to information not restricted by geography or rigid set of friends but rather a flexible and dynamic method to keep up with your many interests and vast imagination.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sy Truong
MetaXceed, Inc. (MXI)
42978 Osgood Rd
Fremont, CA 94539-5627
tel: 510.979.9333
fax: 510.440.8301
E-mail: sy.truong@meta-x.com
Web: www.meta-x.com