Paper 304-2012

Topic Discovery, Tracking, and Characterization of Social Media Conversation for Point of Origin and Dissemination Discovery: Structural Precursors to Topic Determination in Text Corpora

Barry de Ville, Gurpreet S. Bawa, SAS Institute Inc., Cary, NC

ABSTRACT

Social Media conversations cover a wide range of human experience often in novel, sometimes revolutionary ways. The words that are used to express content can reflect mundane events or can sometimes reflect radical new ways of thinking and behavior. The correct identification of conversation in real time is essential in order to effectively respond or intervene in a timely fashion.

This presentation demonstrates the interdependency between vocabulary use and social networks that is critical to understanding conversations in social media channels, where they came from, where they are going and what impact they are likely to have. We show how topics are extracted from social media and how they are identified and tracked. Examples of the language use, social networks, and associated impacts are demonstrated.

INTRODUCTION

Human communication -- like vocal utterances and written expressions -- is, of course, social actions. The social nature of human communication is especially clear in the use of written expression in social media. Unlike many documents, whether published or personal, social media conversations are often multi-sided, interactive and iterative: the conversations are often more like free-flowing cocktail parties than orchestrated staged productions. Hence, the social characteristics of the interaction on social media are important, even primary, determinants that animate the discussions and which drive their development.

The social nature and intensity of social media places a premium on the incorporation of social analytics – in addition to standard text analytics -- in order to effectively interpret and represent the various forms of social communication contained within it. This is not to suggest that text analytics lacks a social dimension: text analytics almost always implicitly incorporate social dimensions in their operation.

The approach outlined here formalizes the social nature of social media expression and specifically includes social community-based units of analysis in the determination of textual meaning and interpretation. This community orientation is useful in textual meaning determination in general. However, our sense is that – especially in the context of social media -- it is a critical pre-requisite in determining the point-of-origin of a given discussion and in the analysis of conversational cascades, explosions, and message dissemination in general. This is because cascades of messages cannot occur without a network structure. While the network has technological enablers, its structure primarily serves a social purpose so is thereby primarily socially determined.

THE ROLE OF COMMUNITY IN SOCIAL MEDIA

Here we outline the role of the social community in text expression and interpretation and in message dissemination. We distinguish various community contexts for social expression, ranging from close personal relationships such as family and intimate friends to primary – and usually participatory -- interest groups, to broader, more geographically and temporally dispersed communities based on group social, political or economic affiliations (for example, a brand-preference interest group or leisure-time activity relationship). Our view on convenient community groupings – that might be overlapping on occasion – is shown in Table 1.

Community Groupings – Ranging in Size, Distribution, Level of Participation and Intimacy
• Broader interest communities. For example, politics, spectator sports, brand affiliation groups • business associates and common interest groups (participatory, for example, cricket league) • close friends and family

Table 1: Community Classification (Overlapping)

Although social community is an important, indeed necessary, construct for the analysis of antecedents that shape meaning and interpretation of written expression, it is only one of many such antecedents or precursors. As described in Appendix A there are many antecedents to a given expression. In many ways the social community captures a number of these and can be considered a strong proxy measure for these other antecedents[1].

THE GRANULARITY OF TEXTUAL EXPRESSION

Language and expression consist of lower order tokens that are often combined or aggregated to form higher-order meanings. This meaning construction methodology is explicit in languages such as Chinese that incorporate root-form characters coupled with other characters to create more complexity and meaning capture as the language evolves. English works this way too: a given document can contain "vote", "ballot box", and "record" + "vote" in various sentences. Here we see the beginnings of a methodology that can incorporate lower level tokens into higher (second and third order) level tokens to express a wider range and generality of meanings.

The SAS® text analytic tools – for example SAS Text Miner® (TM) and SAS Enterprise Content Categorization® (ECC) incorporate these methods of increasing token generality:

In the identification of noun groups: ballot box

In the identification of concept links: record vote

In the identification of topics and categories (found in the ***Topic Node*** and ***Generate subcategories*** functions of Text Miner and ECC respectively)

For our purposes, we consider textual granularity as lying on the rough ordering shown in Table 2.

Textual Granularity
• class • category • topics • collocation • compound • noun group • proper noun • entity • raw

Table 2: Rough Ordering of Textual Expression in Increasing Levels of Generality

[1] In spite of the importance of community membership as discussed herein, community is not always (or even usually) included as a primary unit of analysis in social media analytics. See, for example, Bollen et. al. (2011). This unusual treatment of mood state in social media, which, while quite advanced, gives no consideration to the community context to mood state development and expression.

This ordering is theoretically exhaustive: at the highest level, we use an Integral classification framework (adopted from Wilbur, see Weiner, Irving B. & Craighead, W. Edward, 2010) that will theoretically include any possible written expression.

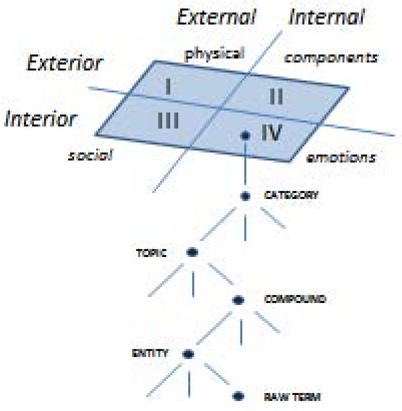


Figure 1: Structure of Highest Level Class (Illustrating Potential Associated Lower Level Mappings)

IMPLICATIONS

When the role of community for expression resolution and description in social media is identified as an analytical priority then a number of implications present themselves for consideration and inclusion in textual expression description and interpretation:

1. Topics of conversation that can be described at the level of the global corpus depend on and are most accurately described as aggregations of topics that are discussed in the various (sometimes overlapping) sub-communities;
2. Different communities refer to the same content in different ways – not only are different words used but the variety and granularity of description varies as well. (Here we refer to granularity as a roughly ordered hierarchy of descriptive meaning that ranges from detailed and specific to more general and all-encompassing). So while one community might refer to “snow”, for example, a skiing-interest group will refer to “sugar”, “corn”, “hard-pack”, “boiler plate” and “powder”, among other descriptors, and might not use the word “snow” at all.
3. An ontological description of the corpus will contain community-based dimensions. Topical and conceptual descriptions will cut across these various dimensions at different levels of generality and at different levels of granularity. So while “Snow” might lie at the very top of the ontology, lower layers will include “sugar”, “corn-snow”, “hard-pack”, “boiler plate” and “powder-snow”. In addition to community dimensions, a full ontology will include multiple levels of generality: raw terms, compound terms, collocated terms, composite terms (including topics), and categories (that, in turn, roll up into classes). Any given community will not usually use all levels of an ontology for a given topic or concept. However, the use or appearance of a term on one level will imply its use or appearance on all related levels. In this way the occurrence of a topic or concept can be inferred on the basis of the structure of the ontology.

ESTIMATION APPROACHES

Estimation and disambiguation have many challenges. Classic rules of inference make the estimation task more manageable, reliable, and accurate:

Temporal order: contents of earlier time-stamped documents and threads precede and therefore influence subsequent documents.

Grain: fine-grained, detailed terms influence the development of higher-level, more general grain terms.

Sequential, temporal, granularity. The estimation model, as illustrated in Appendix A (reproduced here) is ordered left-to-right from the most temporally antecedent, lowest-level term to the highest level, most recent classification.

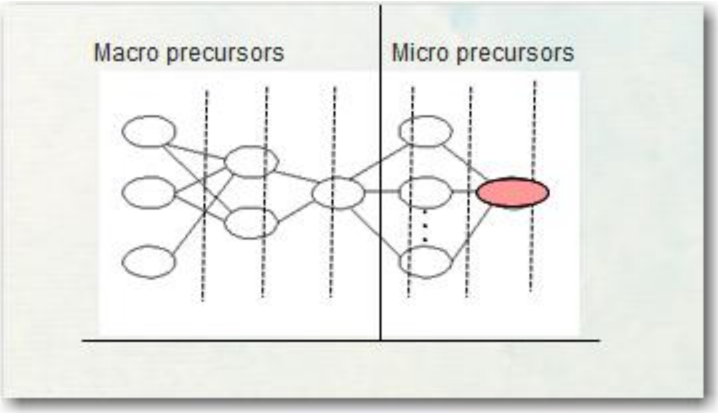


Figure 2: Example of Precedence Relationships in Analytical Model Construction

On the extreme left of Figure 2 we can foresee the inclusion of the operation of macroscopic factors such as network or social role, network social structure (leader, follower, interests, and affiliations). Moving towards the center portion of this semi-causal sequence of events we might see the operation of such individual factors as mood, personality, and gender. Finally, at a micro level we might expect to observer time-of-day features, for example, or perhaps word sense disambiguation based on contextual referents. We would see this latter process in the disambiguation of co-references (anaphora) such that “he”, “she” or “it” would refer back to a previous message on the thread.

Token Specificity

Tokens – as the various granularities of word combinations are referred to here – resolve from the lowest (raw term or root form) level to the highest class level.

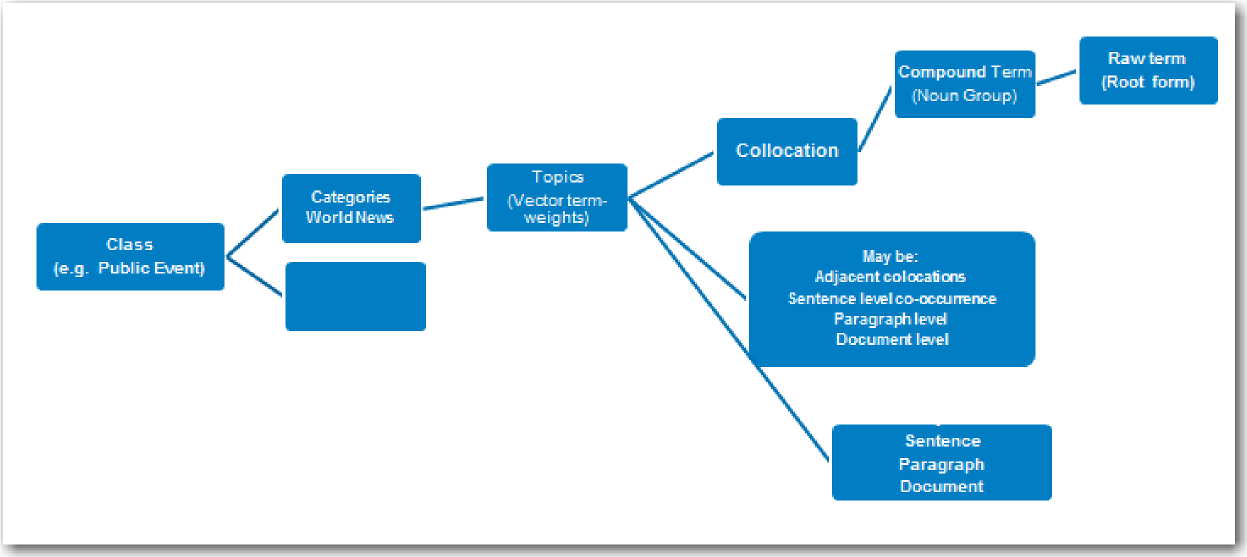


Figure 3: Written Word Granularity as Precedence Relationships

Community Specificity

Term sequences and generality resolve first in the most immediate community, next in the more distant community and finally in the most distant community (and so on). Community distance is illustrated below.

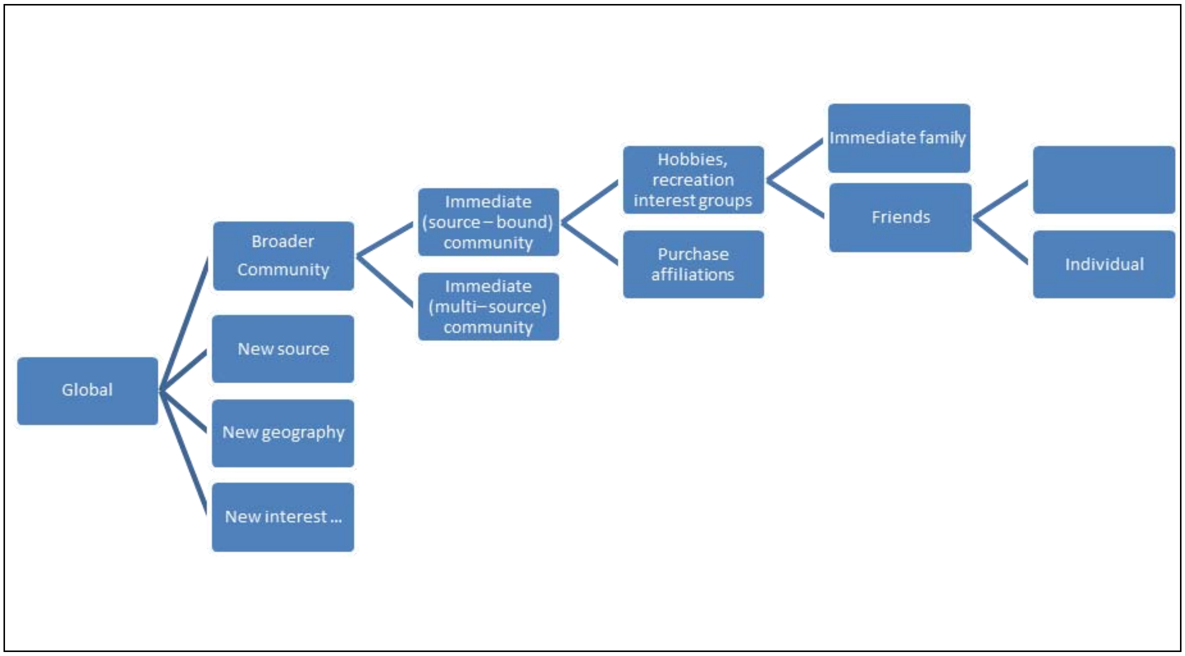


Figure 4: Community Levels of Granularity as Precedence Relationships

MEMES, MIMES AND PUTNAM'S THEORUM

Language evolves at a breathtaking rate. According to Richard Alleyne (Alleyne, 2012) researchers at Harvard University and Google found that English was expanding by 8,500 words a year in the new millennium and now stands at 1,022,000 words. Most dictionaries list between 70,000 – 125,000 words According to the web site "The Second Edition of the 20-volume *Oxford English Dictionary* contains full entries for 171,476 words in current use".

Much of new language usage – and many of the new terms – spring from the operation of memetic and mimetic factors. Memes, as defined by Brodie (Brodie, 1996) are "… unit(s) of information in a mind whose existence influences events such that more copies of itself get created in other minds".

Figure 5 shows the operation of memes in research carried out by Leskovec, Backstrom, Kleinberg. (Leskovec et. al., 2009). Because of their unique method of analysis they were able to identify that "lipstick on a pig" and "fish in a new wrapper" were memes (both expressions refer to failed attempts to disguise the underlying object by applying superficial "cosmetic" fixes.

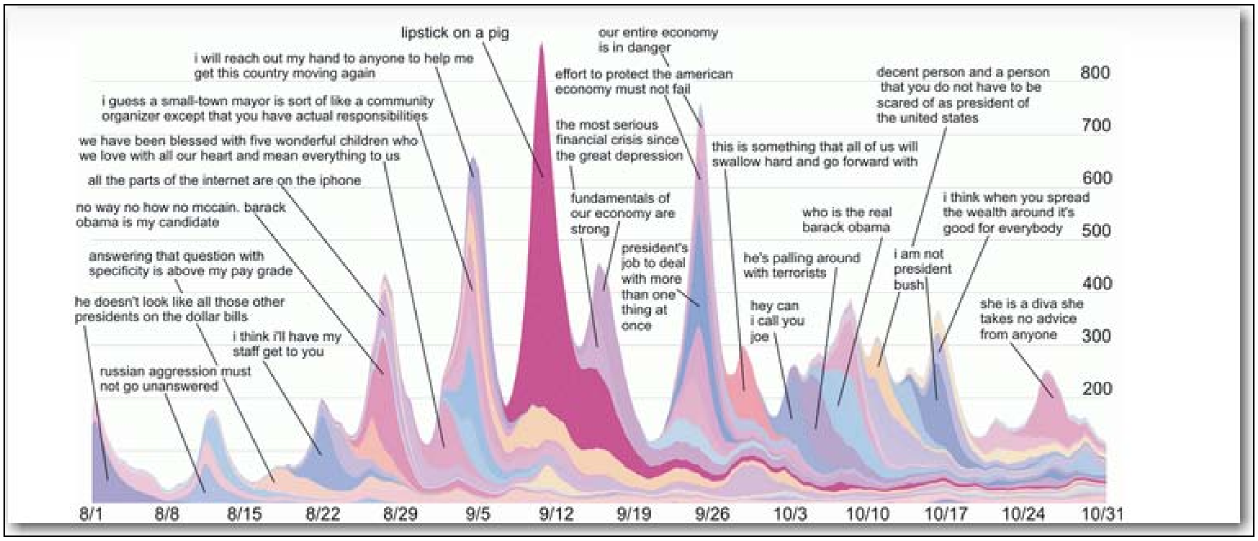


Figure 5: Evolution of Memes in Social Media (from memetracker.org)

A robust search and projection approach to topics, themes and so on in social media must account for such near-match changes in expression.

Just as humans employ memes as language forms that have higher survival value so too do humans employ mimes as survival strategies. As the saying goes “imitation is the sincerest form of flattery”. If something works, humans copy it. With the maturation of the internet entire industries have evolved that specialize in harvesting intellectual property … through both legal and illegal means.

The reinforcing nature of memes and mimes – new forms then the mass copying of new forms – is a pervasive mechanism underneath all contagion estimation approaches in social media. Finding near-matches to new forms is a formidable challenge.

Putnam’s theorem states that there is “… no unique mapping between words and the world: even if we know the conditions under which sentences are true, we cannot fix the way their terms refer”. (quoted in Varela et. al., 1993).

The net effect of Putnam’s theorem – not to mention the many other threads that confirm our darkest suspicions -- is that word meaning is forever and always determined by *social convention* (and ***only*** social convention). Words like “pious”, for example, which might have formerly carried a positive connotation now appear negative. Many other examples of meaning changes abound.

Because language is constantly-changing, this is bad news for dictionary lookup methods for word sense disambiguation. At least in social media, however, this is somewhat good news: because word sense is derived from social convention then we will find that different social contexts reveal different word meanings.

Social media is an ideal medium for exploring and exposing social contexts. A strategy of identifying word meaning through the operation of factors involved in a social contexts, as illustrated in Figure 2, is expected to yield rich dividends as we track the source and likely evolution of a range of topics discussed in social media.

COMMUNITY IDENTIFICATION

The community context of any given act or expression can be defined by the 4 W’s: who (with), where, when, what (why and how)?

The **who (with)** of a social communication is normally defined by threads or addresses. All open forums and fan pages contain exhaustive thread indicators that keep track of which messages go to which recipients. In Twitter, the theoretical community includes followers – who cannot be tracked – while the actual (reduced-size) community includes replies (direct tweets), mentions and re-tweets.

The **where** can be resolved by identifying the platform – Twitter, Facebook, and so on. The physical location can often be inferred by the registration information of the user, is sometimes given by geo-coordinates, and can sometimes be inferred by locations indicated in the message itself.

The **when** can be extracted from the time stamp of the message or can be more broadly determined by time and place referents that appear in the message.

The **wha**t is the information that is resolved through the operation of the ontology. There is an element of bi-directionality here, however, since as defined above, the ontology itself depends on the content that is being discussed within the context of the community.

The **why and how** (talk about word change! … we now have 5 W's – and an H) are important as well. **Intention** may well be the single most important determinant of word meaning. A lot of intention can be extracted from the social context – for example, a purchase is a purchase intention, and so on. Much of the how – via query, via post, via request for comment, and so on – can also be determined.

For purposes of this initial exploration community can be derived in the following fashion:

- The method implemented in SAS® PROC Optgraph based on an algorithm proposed by M. E. J. Newman (2011).
- Semantic determination (for example, clustering of SVD factors).
- Followers grouping (form groups based on who is followed, referenced and so on)

Multiple community composites – friends, family, peers, interest groups, larger communities – can be derived through the consolidation of 1 or more of these methods[2].

DISCOVERY

No one word in any language has an unchanging, fixed sense in a given context. Language usage is constantly changing in social media. Moreover, new words are constantly invented. So no fixed dictionary will ever provide a full-service, one-stop shop for interpreting the meaning of text. Discovery is always required.

Even in the simplest case where individuals use semantic tagging to identify word meaning it is useful to have community affiliation. Two individuals might well identify two separate meanings for the same word. In this case knowledge of community and the social tendencies associated with the community will help resolve the conflicting meaning.

Another form of discovery involves the identification of synonyms and uses textual cues such as "alike", "such as", "the following" and so on. Dictionary entries can also be searched and the most popular meaning found can be assigned to give meaning to the target text.

Conditional statistical distributions are used in a variety of linguistic applications. The SAS® Text Miner® PROC TM Belief uses this approach to identify high probability terms given the occurrence of one or more other terms.

A more general form of discovery involves the identification of collocations such that unique combinations take on an unambiguous meaning precisely because of their uniqueness. Latent Dirichlet and other latent structure approaches (such as the SVD and Topic Factorization used by SAS®) fall into this class of techniques. The concept link diagram in Text Miner is another such approach. In all cases conditioning the results by community proves helpful: since the vocabulary that is used is community-based, variability between communities will always be greater than variability within communities and therefore sense disambiguation between communities will be facilitated.

MECHANISMS

Based on the discovery methods described above, we have constructed various mechanisms to generate tokens at various levels of granularity, as described below. Compute topic growth model. This process will find a topic and demonstrate topic growth over time (by thread if possible). It also shows the drivers of the growth.

[2] We only include numbers with less than 150 relations to other nodes. "*Dunbar's number is a theoretical cognitive limit to the number of people with whom one can maintain stable social relationships. These are relationships in which an individual knows who each person is, and how each person relates to every other person.*"

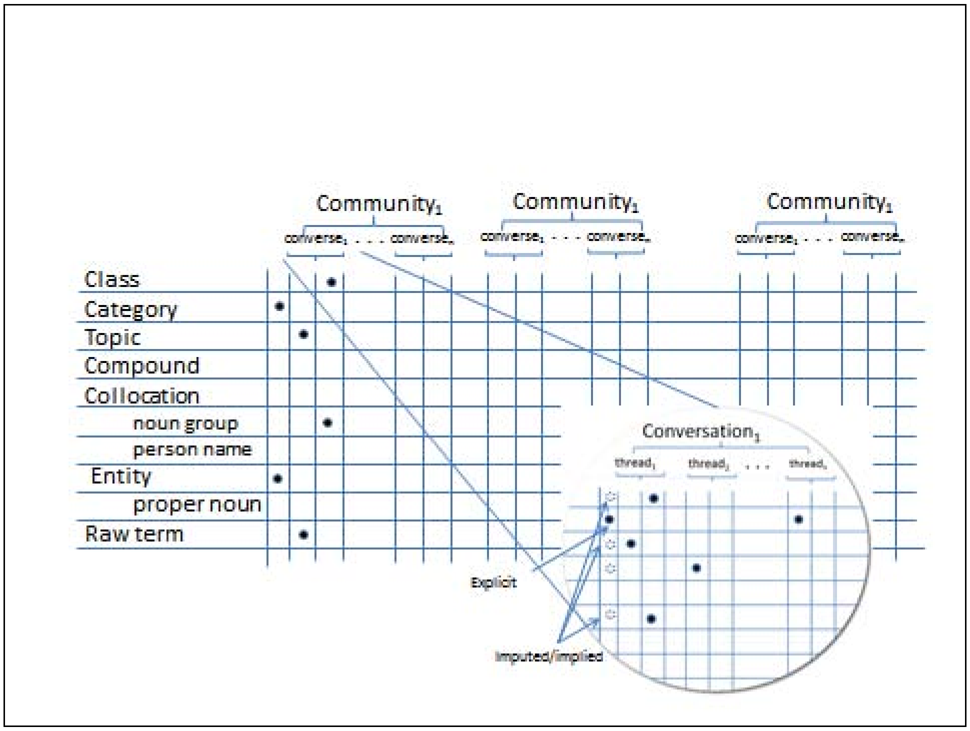


Figure 6: Example of Community and Linguistic Granularities

1. Since granularity and content can be considered community attributes, then documents can be grouped by community. This means that a higher level category can be distributed down and across to all the lower level "grain attributes" of all the members of the community. Similarly, lower level granularity occurrences can be promoted up and across. This provides a mechanism to "fill in the blanks" and to produce a full ontology on a community basis. (In cases where there are more candidates than positions – for example, competing categories – a voting strategy can be devised to arbitrate the appropriate meaning).
2. Roll up tokens by Grain by Community (that is, compress threads)
3. Roll up tokens by Grain (Corpus Level, no community indicator)
4. Fill in blank spots by thread (if token is in one thread, it is in all threads). Move from most distant to most proximate. Store as estimated value.
5. Roll up by Grain by Community.
6. Compute phrase or term point of Origin. Any row or granularity on the ontology can be used to map to any higher or lower point in the ontology. This allows for "fuzzy similarity mapping" in the retrieval of topic origins and also constructs memetic synonyms (http://en.wikipedia.org/wiki/Memetics) for the term entries and various levels of granularities

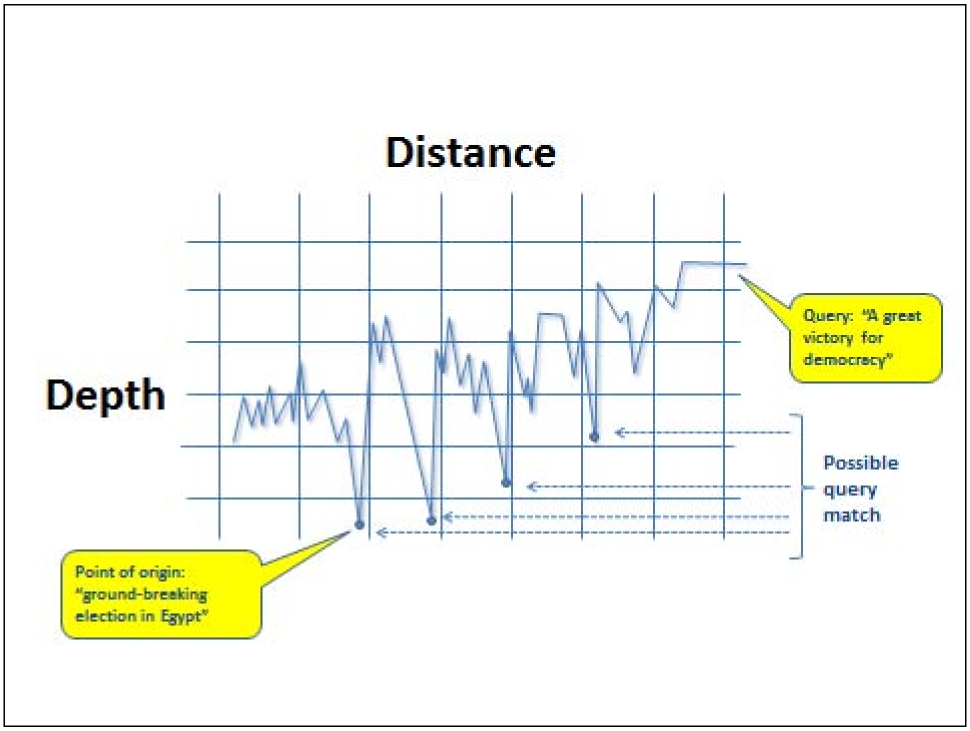

Distance
Depth
Query: "A great victory for democracy"
Possible query match
Point of origin: "ground-breaking election in Egypt"

Figure 7: Example Search Trace for Point of Origin Analysis

CONCLUSION

The social dimension of written communication is operationalized as an explicit method for word sense inference and disambiguation by incorporating community as an instrument of textual interpretation. Given two competing meanings for a particular written expression, we choose the meaning that is closest to the most specific and intimate community group that is implied by the expression.

Just as we take advantage of precedence relationships as expressed through community structure, we also take advantage of precedence in token generality by imposing a rough ordering to written expression that ranges from the raw or root form through various stages of increasing composition and complexity up to the highest level of category or class.

These two mechanisms – community and precedence – enable us to infer meanings and relationships between various written expressions as an ontology based on community. This ontology provides a mechanism to equate token meanings, assign synonyms and memetic variations. This provides a flexible and powerful mechanism for tracking the origins of a given expression across various social media sources through time. It also provides a similarly flexible mechanism for tracking and predicting cascades and contagion, along with the associated driver mechanisms that appear to be influencing the discussion.

REFERENCES

Alleyne, Richard. 2012. “English language has doubled in size in the last century.” *The Telegraph (http://www.telegraph.co.uk/)*, Tuesday, 13 March.

Ball, Brian, Karrer, Brian, and M. E. J. Newman*.* 2011 “An efficient and principled method for detecting communities in networks.” *Phys. Rev*. *E* 84, 036103.

Bollen, Johan, Mao, Huina, and Xiaojun Zeng. March 2011. “Twitter mood predicts the stock market.” *Journal of Computational Science* 2(1):1-8.

Brodie, Richard. 1996. *Virus of Mind*. Hay House, Inc. (http://www.hayhouse.com/).

Leskovec, J., L. Backstrom, J. Kleinberg. 2009. “Meme-tracking and the Dynamics of the News Cycle”. *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2009.

Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press. As quoted in Varela, F., E.

Thompson, E. Rosch. 1993. *The Embodied Mind*. MIT Press. Page 233.

Weiner, Irving B., and W. Edward Craighead, eds. 2010. "Integral Psychology." In *The Corsini encyclopedia of psychology,* 2(4):830. New York: Wiley.

ACKNOWLEDGMENTS

None of this work would have been possible without the help of many others, primarily:

James Cox, Hector Cobo, Mauricio Salusti, Russell Albright, Saratendu Sethi, Richard Foley

CONTACT INFORMATION

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

Barry de Ville
SAS Campus Drive
SAS Institute Inc.
E-mail: Barry.deVille@sas.com

Gurpreet Singh Bawa
SAS Campus Drive
SAS Institute Inc.
E-mail: Gurpreet.Bawa@sas.com

APPENDIX A

THE CONTEXT OF WRITTEN EXPRESSION

The communications context is illustrated Figure 1. Here we see that the situational context of any given utterance or expression lies in the context of a sequence of situational factors. All of which tend to constrain or shape the form and content of the expression. The expression itself is shaped further by follow-on situational factors which also influence the form and meaning of the expression. As a result of this broad context NLP approaches have evolved to incorporate more information so as to render the influence of context and so more effectively extract meaning from expression.

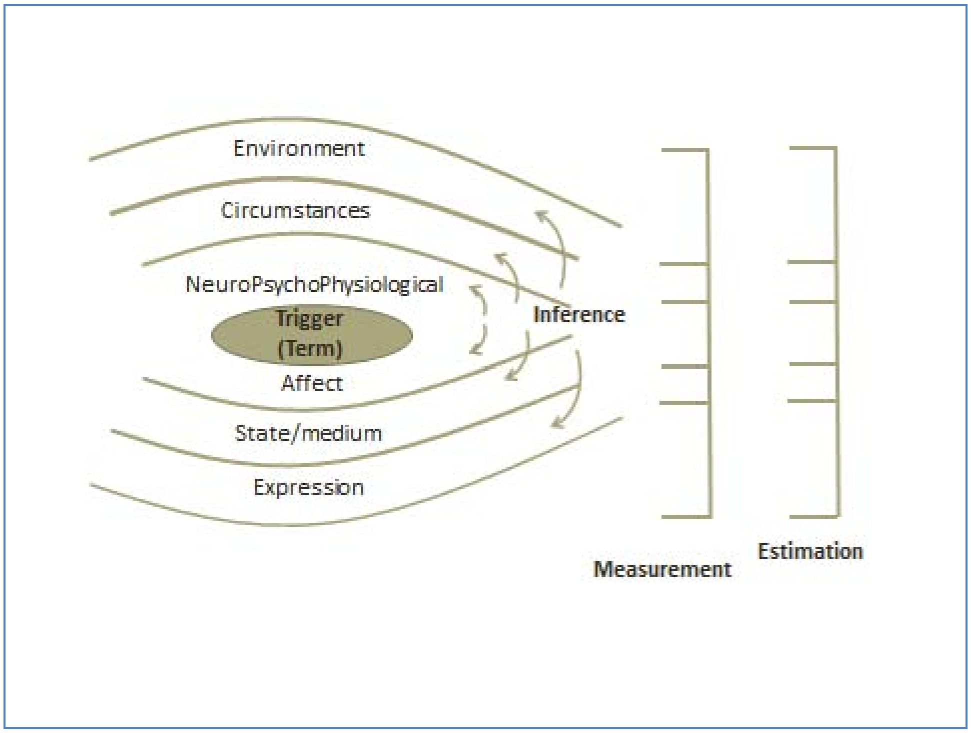


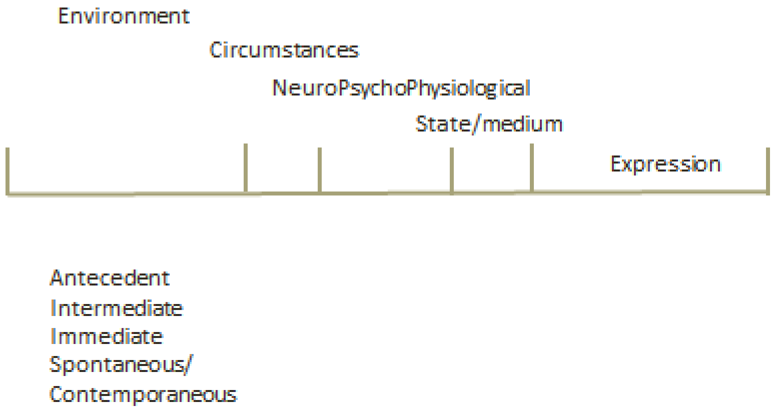
Figure 1 illustrates the framework that gives rise to an emotional impression and resulting expression. It presents an overlay of the 3 major ways in which aspects of the framework are tracked and interpreted in order to derive meaning from the expression, usually captured as a term (or sequence of terms) in a document.

The framework reads from top to bottom and tells us that emotional expressions emerge from an environment and situational context. This constitutes the total outer layer of events surrounding the human agent. Within the human agent, we have neurological, psychological and physiological processes that receive information from the environment and transform it into an emotional response.

The inner response boundary of the human agent includes a set of internal processes that create an effect. In the outside layer of responses lay the particular style and medium of delivery and finally the formation of the emotional expression itself.

Illustrated to the right of the diagram are the main methods that used to derive meaning from this sequencing and interplay of processes: inference, measurement, and estimation. Explicitly or implicitly the terms that are found in a document attempt to infer the internal and external processes that trigger the human agent's reactions and responses. Ideally these inferences are accompanied by measurements that are specifically directed at the external sources of emotional triggering and both internal and external emotional expression formation and processing.

In cases where measurement is not exact and where inference is difficult a variety of estimation techniques are available. These include simulation, table lookups, and quantitative modeling.



In reality, whenever we view a given expression (or utterance) it is the result of a pre-existing sequence of events that stretches out into the past. In practice, we have a number of indicators or measurements of the sequence of events that occur prior to the recording of the particular expression. And we know with certainty that these sequences of events have a profound effect on the form of the expression. Nobody questions this: expression does not emerge from a vacuum.

In fact, we know that the particular kind and sequence of events have a strong effect on the form and content of the expression. We also know that similar expressions result from different – sometimes radically different – sequences of precursors. The range of expression is always smaller than the range of circumstances that gave rise to it. So expression will always involve an element of uncertainty.

It is axiomatic that there is never any single, permanent equivalent meaning for any given expression.

When we parse the meaning of a given expression – say, as an example, a written sentence – we are always trying to infer the meaning of the sequence of events that gave rise to the expression. Carried out in isolation, this is an impossible task. No amount of even incredibly deep parsing will ever allow us to infer the unobserved series of circumstances that preceded a given event.

With empirical data observation and methods this limitation of parsing can be easily overcome. If we want to know about the sequence of events that preceded a given expression, we can often discover a lot of things. Many of the things that we discover will have a strong – even determining – effect on the characteristics of the expression.