VOXSUP: A Social Engagement Framework

Yusheng Xie[1], Daniel Honbo[1], Kunpeng Zhang[2], Yu Cheng[2], Ankit Agrawal[2], Alok Choudhary[1]
[1]Voxsup Inc.
{yves,dan,alok}@voxsupinc.com
[2]Northwestern University
{kzh980,ych133,ankitag}@eecs.northwestern.edu

ABSTRACT
Social media websites are currently central hubs on the Internet. Major online social media platforms are not only places for individual users to socialize but are increasingly more important as channels for companies to advertise, public figures to engage, etc. In order to optimize such advertising and engaging efforts, there is an emerging challenge for knowledge discovery on today's Internet. The goal of knowledge discovery is to understand the entire online social landscape instead of merely summarizing the statistics. To answer this challenge, we have created VOXSUP as a unified social engagement framework. Unlike most existing tools, VOXSUP not only aggregates and filters social data from the Internet, but also provides what we call Voxsupian Knowledge Discovery (VKD). VKD consists of an almost human-level understanding of social conversations at any level of granularity from a single comment sentiment to multi-lingual inter-platform user demographics. Here we describe the technologies that are crucial to VKD, and subsequently go beyond experimental verification and present case studies from our live VOXSUP system.

Categories and Subject Descriptors
H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval – *Information filtering, Selection process*

General Terms
Algorithms, Experimentation.

Keywords
Opinion mining, Topic model, Social ranking.

1. INTRODUCTION
Online social media websites are currently among the most popular Internet applications. Hundreds of millions of Internet users go to websites like Facebook and Twitter several times a day. Naturally, the massive use of public social media has become valuable for scientific research due to the large amount of human-generated contents. Social media can be used as datasets and attract a lot of academic interests ranging from large-scale complex network modeling to individual behavioral targeting [1]. Previous works dealing with social media data return impressive results in many well-defined problems. For example, sentence-level sentiment analysis done by [2] is used in real applications. Such research efforts help us to realize how valuable social media contents can be from many different viewpoints. However, since every major social networking service (e.g. Facebook, Twitter, LinkedIn, etc.) avidly supports Google-style advertising, just solving isolated problems is definitely not the only way researchers can interact with social media. Driven by commercial marketing interests, the addition of knowledge discovery to social media analysis will greatly increase product-marketing performance. The increase of performance does not emerge from a list of individual solutions like sentiment analysis; it has to come from a unified framework. Social media analysis needs such a framework.

VOXSUP is our answer to this need. VOXSUP exceeds the performance of most existing tools because it not only aggregates and filters social data from the Internet, but also performs what we call Voxsupian Knowledge Discovery (VKD). VKD provides an almost human-level understanding of social conversations at any level of granularity from a single comment sentiment to multi-lingual inter-platform user demographics. In the following sections, we describe the technologies that are crucial to VKD.

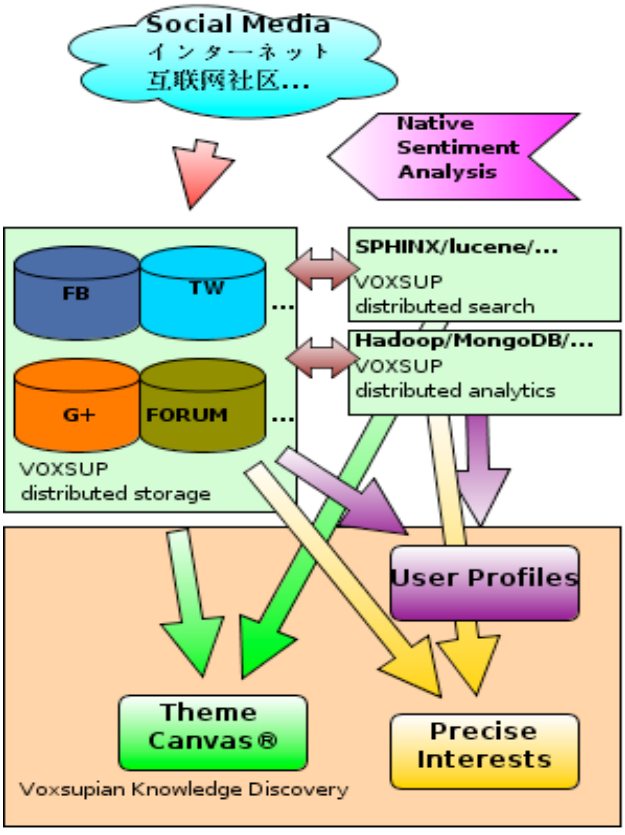

Figure 1. Overview of VOXSUP architecture.

KDD'12, August 12–16, 2012, Beijing, China.

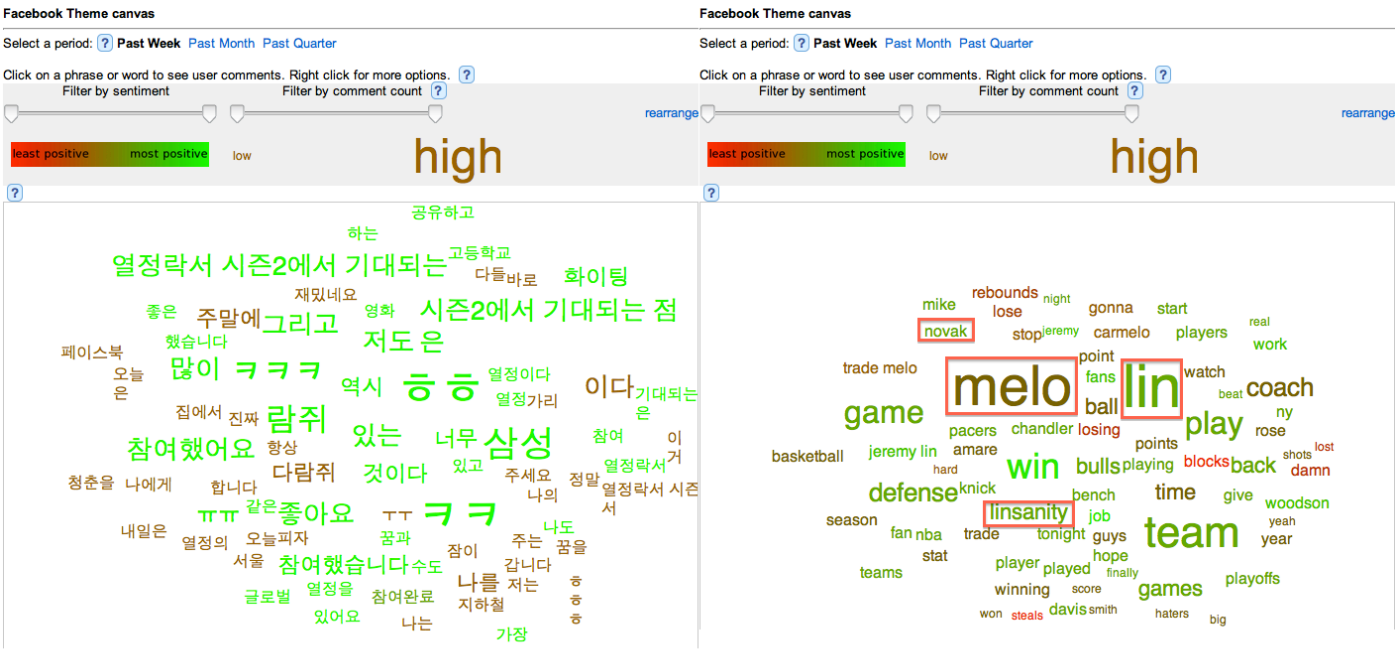

Figure 2. Theme Canvas® screenshots: (a) Samsung results showing multi-language capability. (b) New York Knicks results showing the quality by discovering frequently talked about Jeremy Lin, *Linsanity*, etc. Note that the results are fully tunable by changing the time period, sentiment, frequency filters and other parameters.

2. SYSTEM DESCRIPTION

VOXSUP architecture is shown in Figure 1. In short, VOXSUP monitors multi-lingual social media content, performs sentiment analysis on this content, pushes it to scalable infrastructure, and finally extracts the human-level knowledge from this content.

2.1 Media Monitored

VOXSUP monitors multiple channels of social media in more than 7 languages, these channels include over 10,000 of the most popular Facebook public walls, over 3,000 Twitter handles, over 800 YouTube Channels, over 2,000 fo the most popular Google+ public pages, over 1,500 LinkedIn company profiles, over 500 product categories with customer reviews from Amazon.com and over 300 active blog sites and online forums. For most of these sources, we have all its data to date since 2009 or its inception. In VOXSUP, we try to keep as much information as possible without violating any regulations. For example, for each Facebook wall, we keep all public comments, "likes", and public user profiles. Despite VOXSUP's enormous capacity, we can still keep our data synchronized with the providing media in soft real time.

2.2 Sentiment Analysis

For sentiment analysis, we use our state-of-the-art Sentiment Elicitation System (SES). Complete details can be found in [2].

Algorithm 1: CSR Compositional rule.

Input: *arg1*: sentimental word; *arg2*: sentimental word;
Output: *sentiment*: sentiment for the composition;

```
1   If arg1 is negative
2     If arg2 is not neutral {return S(arg2)}
3     Else {return -1}
4     End if
5   Else if arg1 is positive and arg2 is not neutral {return S(arg2)}
6   Else if S(arg1) == S(arg2) {return 2*S(arg1)}
7   Else if (arg1 is positive and arg2 is neutral) or
8       (arg2 is positive and arg1 is neutral) {return S(arg1)+S(arg2)}
9   Else {return 0}
10  End if
```

Compositional Semantic Rule (CSR) estimates the sentiment from a grammatical point of view. The algorithm looks for 12 specific patterns in the text, two of which we give as examples in Table 1 and our Compose function is described in Algorithm 1. Currently, we only apply this method to the English language.

Table 1. Two examples of CSR rules

Rules	Example
S([adj] to [verb])=Compose(adj,verb)	*Unlikely* to *harm* the Earth
S([noun] be [adj])=Compose(noun,adj)	*Damage* is *minimal*

Numeric Sentiment Identification (NSI) estimates the sentiment in a numeric scale, which provides a convenient feature in the later learning stage. The idea is to use a training set with labeled sentiment to assign a numeric score to each of the words in the training sentences; then assign a score to any new sentences based on the scores on words. We assign a score to each word by

$$\text{Score}(w) = \frac{\sum_{i \in P} (n_{r_i}/N) r_i f_{r_i}}{\sum_{r_i} (n_{r_i}/N) f_{r_i}}, \quad (1)$$

where P represents the set of appearances of word w, r_i represents the associated label in each appearance of w, N is the number of entities in our training data set, and f_{r_i} is the number of entities with label r_i. After obtaining a score for each word, the score for sentence is calculated by a weighted linear combination. Note that in our original work [2] we distinguish between adjectives and adverbs, but the basic idea is the same.

Bag-of-word and Rule-based (BR) is specially designed for social media type of text where people type a lot of nonstandard phrases like Internet sarcasm, emoticons (e.g. :)), acronyms (e.g. "lol" for "laughing out loud"), etc.

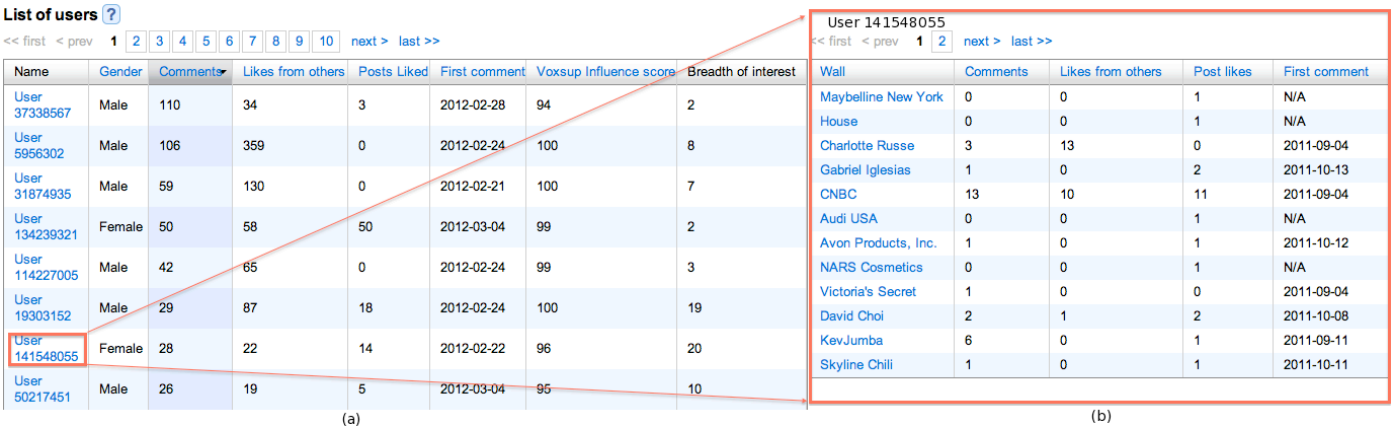

Figure 3. User profile screenshots: (a) Partial list of the social users associated with the scenario *New York Knicks*. (b) More details about this particular user showing her activities on other subjects and interests. Note that all the statistics shown are contingent upon parameters (e.g. time window) chosen.

The final sentiment is obtained by combining the three results through a random forest model.

2.3 Distributed Solution for Big Data

VOXSUP is designed to be a cutting edge data mining system that is (almost) always available, highly durable, and easily scalable. Different from many data mining projects, which first receive a massive, static dataset and then perform mining tasks on it, VOXSUP contains a cluster of several transactional databases and high-dimensional data warehouses. Each social channel VOXSUP monitors, e.g. Facebook, has an individual transactional database that has to handle soft real time updates as the source channel updates. While enabling this feature for just the 10,000 walls Facebook is not trivial, VOXSUP does it for almost all major social media. What is more challenging is how to make sense of this massive amount of data from so many different yet correlated sources. The individual transactional database for each source in VOXSUP scales easily because such databases are independent of one another. However, this independency vanishes when we build the data warehouse. For example, an end-user may be interested in knowing what the hot topic talked about by certain group of social users is; he/she wants to know this for the entire social landscape. Traditional independent analytics on individual platforms will be, at best, biased estimates. VOXSUP can discover knowledge in real time from cross-platform heterogeneous information and therefore give end-users timely and unbiased insights.

To support demanding data mining tasks like the above example, we enable free text search on all of the text in our system through multiple scalable indexing / searching solutions.

2.4 Voxsupian Knowledge Discovery

VKD provides comprehensive, unbiased, actionable insights of the entire social topology through its 3 major components: Theme Canvas® discovery, user profile discovery, and precise interests discovery.

Theme Canvas® discovers the most talked about keywords and phrases for any given social channel in any given time interval. Different from traditional "word cloud" applications, Theme Canvas® encodes general social opinions associated with each term in gradients of two colors, and represents the scope or influence of each topic / term by its font size on canvas; clicking on any of the phrase takes the end-user to all of the original sentences/paragraphs that mentions the chosen phrase. We find the sentiment of each word (phrase) S_w on the canvas by latent topic models [3]:

$$S_w \propto \sum_{\substack{z\in \text{all} \\ \text{topics}}} \sum_{\substack{d\in \text{all} \\ \text{documents}}} S_d \cdot p(w|d,z), \tag{2}$$

where S_d is the sentiment for each document (e.g. comment, tweet) determined by SES. In addition, Theme Canvas® incorporates streaming data and returns results in real time. We achieve this by methods from item counting in streaming data [4]. Figure 2 visualizes this application.

User profile discovery in VKD analyzes each social user's public interests network. A Voxsupian user profile is the summary of this social user's presence and activities on different public pages across all platforms over a period of time. Like many available social profiling tools such as Klout, our user profile ranks the most influential users. But unlike most other social user profiling tools, the Voxsupian user profile distinguishes scenarios. For example, the online accounts for politicians and rock stars are influential and popular; but without distinguishing scenarios, their popularity does not translate into any useful information for, for example, a local business owner who wants to maximize viral reach by micro-targeting local residents who are influential for her brand and business. VOXSUP generates a customized user profile for each scenario. In the mathematical analogy, if a traditional user profile is represented as a vector of homogenous objects: $U=[o_1,o_2,...,o_n]$, a Voxsupian user profile is essentially a tensor: $\tilde{U}=[T^{(o_1)},T^{(o_1)},...,T^{(o_n)}]$. Tensors are confusing to many people, so we summarize all this information in each Voxsupian user profile into a single scenario-dependent measure, the Voxsup Influence Score, as an indicator of user's scenario-dependent influence power. To calculate this score, we map $\tilde{U}$, a Voxsupian user profile, to a score vector $SC\in\Re^n$. For each scenario $i\in\{1,...,n\}$ in $\tilde{U}$, let $A_i=\{a_1^i,...,a_m^i\}$ be his/her set of activities, $S_i=\{s_1^i,...,s_m^i\}, s\in\Re$ be the corresponding set of sentiment representations, and $P_i=\{p_1^i,...,p_m^i\}, p\in\Re^+$ be the corresponding set of popularity measures. Then we define the Voxsup Influence Score in an analogous fashion to that of the H-index [6]. For each scenario $i\in\{1,...,n\}$, we define $SC[i]=\tilde{s}\tilde{p}$ where

$$\tilde{s}=\sup_{|s_j^i|}\left\{\left|s_j^i\right|<\left|\left\{s_k^i\in S_i:\left|s_k^i\right|\geq\left|s_j^i\right|\right\}\right|\right\}, \tag{3}$$

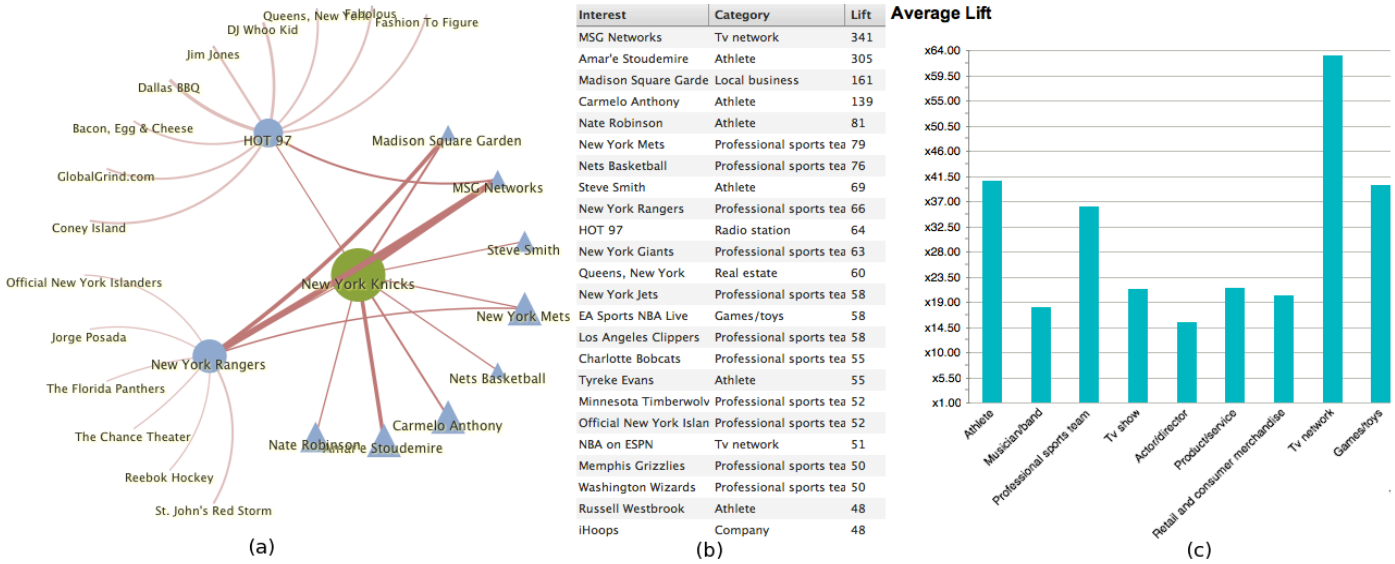

Figure 4. Precise interests screenshots: (a) Visualization of the interests for the description *New York Knicks*; thickness of the edge represents the strength of the connection; size of the geometry of each node represents the predictive confidence; each interest node is clickable and expandable (e.g. *HOT 97* and *New York Rangers*). (b) List view of the precise interests in (a) with assigned category and relative lift. (c) Category-level roll-up lift chart.

$$\text{and } \tilde{p} = \sup_{p_j^i} \left\{ p_j^i < \left| \left\{ p_k^i \in P_i : p_k^i \geq p_j^i \right\} \right| \right\}. \quad (4)$$

Just as TIME magazine ranks the most influential people on both positive and negative impact, in the Voxsup Influence Score we take the absolute value of the sentiment to accommodate such consideration. We empirically approximate the effect of sentimental strength by scalar multiplication. An alternative approach in our practice is to use power scale: $SC[i] = \tilde{p}^s$. The scores shown in Figure 3 are normalized between 0 and 100. Figure 3 contains some examples of Voxsup Influence Score.

Precise interests discovery analyzes the entire public social web and extracts the most relevant/interesting items to a given description. A description here means a description of an interest; it could be a brand, a phrase, a group of users, etc. For example, the brand "Mercedes-Benz" is a description of an interest in luxury cars. VOXSUP system would then query its social database and returns a list of relevant interests that share a similar demographic as the chosen interest. Voxsupian precise interests are particularly useful for the behavioral targeting world. The screenshots in Figure 4 can best illustrate the Voxsupian precise interest. Figure 4 (a) and (b) show two different ways to explore the interests: breadth first and depth first. In the interests graph, we apply the same algorithm to each node interest to get a list of interests and sort the output by "lift". This is an atomic operation, which we call *probe*. Each *probe* terminates on a user-set lift threshold. By setting the threshold to just the interest with the highest lift, *probe* becomes depth first; otherwise, it is breadth first. Suppose X is a random variable denoting average social user's interest distribution and if we are given two interest descriptions x_1 and x_2, we define x_1's lift on x_2 as the ratio: $\Pr(X = x_1 | x_2) / \Pr(X = x_1)$. Figure 4 (c) further shows the lift numbers on category level. VOXSUP assigns categories of interests by clustering the interests based on their online demographics [5].

3. CONCLUSION AND FUTURE WORK

In this work, we introduced VOXSUP as a social engagement framework. To support its large-scale data-mining tasks, VOXSUP employs distributed solutions to handle and deliver its enormous magnitude of data. Based on our previous multi-lingual sentiment identification algorithms, we present the Voxsupian Knowledge Discovery as three major components: Theme Canvas® discovery, user profile discovery, and precise interests discovery. We describe the technical aspects and potential applications of the three components. Future research on network sampling techniques will be very helpful in our data mining practice if such techniques can promise good estimates of the original data using much less computing resources.

4. ACKNOWLEDGEMENTS

This research was supported by Voxsup, Inc.

5. REFERENCES

[1] S. Kim, T. Qin, H. Yu, and T.-Y. Liu. Advertiser-centric approach to understand user click behavior in sponsored search. In *CIKM '11*. ACM, 2011

[2] K. Zhang, Y. Cheng, Y. Xie, A. Agrawal, D. Palsetia, K. Lee, and A. Choudhary. SES: Sentiment Elicitation System for Social Media Data, *ICDM-SENTIRE '11*. IEEE, 2011

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML '06*. ACM, 2006

[4] G. Cormode and M. Hadjieleftheriou. Finding frequent items in data streams. In *VLDB*. ACM, 2008

[5] A. Lancichinetti and S Fortunato1. Community detection algorithms: A comparative analysis. *Phys. Rev.* E 80, 056117. 2009

[6] J. E. Hirsch. An index to quantify an individual's scientific research output. In *PNAS* 102 (46). 2005