Towards a Set Theoretical Approach to Big Data Analytics

Raghava Rao Mukkamala
Postdoc, Software and Systems Section
IT University of Copenhagen, Denmark

Joint work with
Ravi Vatrapu
Professor, Department of IT Management
Abid Hussain
PhD student, Department of IT Management
Copenhagen Business School, Denmark

IEEE BigData 2014, Alaska, USA
July 01, 2014
Road Map

• Part 1: Introduction and Motivation
• Part 2: Social Data Model
• Part 3: Social Data Analytics Example: H & M Company
• Part 4: Conclusion & Future work
Social Media as Business Platform

- Most used platforms: Facebook, Twitter, Youtube, LinkedIn
- Content published worldwide by million of social media users
- Social Data Sets contain valuable information
- Can provide meaningful facts and actionable insights


Facebook-Social Media Platform

- A Facebook wall offers
  - a new and more informal, but a public means for users
  - provision to enter into discussions and debate with other users
  - engagement can range from interactions under normal conditions to activities under extra-ordinary events such as Arab Spring
- Facebook gives you friends, while Twitter gives you followers!
- Structure and Data availability
  - Twitter: Simple and public
  - Facebook: Complicated and Restricted privacy settings
- Research wise: Mostly on Twitter, very less on Facebook

Social Data Research Approaches

• Ethnographical Approaches
  • social meaning inferred from content
  • usually qualitative techniques based on triangulation methods

• Statistical Approaches
  • Computational supported statistical approach (correlation, regression, etc)
  • e.g. study of twitter usage in natural disasters

• Computational Approaches
  • Computational Social Science: interdisciplinary approach, social scientists + computer scientists + mathematicians
  • building models, methods and concepts for analysis of large volume of data

Our Research Approach

• Formal Methods to develop advanced data analysis techniques for social data

• Formal methods: technique to model complex phenomena as mathematical entities
  • abstract, precise and complete

• Current techniques limited to Social Network Analysis based on graph-theoretical approach (Relational Sociology)

• Our approach: based on Set theory and Fuzzy Logics (Associational Sociology)

• Primarily focussed on Facebook data
Research Methodology

- Using Integrated Modeling approach

- Conceptual Model of Social Data

- Formal Model based on Set Theory

- Social Data Analytics Tool (SODATO)
Social Data Conceptual Model

- Social Graph Analysis: structure of relationships emerging from social media use
  - Which actors involved?
  - What actions they perform?
  - What activities they undertake?
  - What artifacts they create and interact with?

Social Data Conceptual Model

- Social Text Analysis: substantive nature of the interactions
  - How the topics are discussed?
  - Which keywords appear?
  - Which pronouns are used?
  - How are far positive/negative sentiments expressed?

Formal Model - Social Data

**Definition.** We define $\mathbb{R}$ as a set of all artifact types as $\mathbb{R} = \{\text{status, comment, link, photo, video}\}$.

**Definition.** We define $A_{CT}$ as a set of actions that can be performed as $A_{CT} = \{\text{post, comment, share, like, tagging}\}$.

**Definition.** Formally, Social Data is defined as a tuple $S = (G, T)$ where

(i) $G$ is the social graph representing the structural aspects of social data

(ii) $T$ is the social text representing the content of social data
Definition. The Social Graph is defined as a tuple \( G = (U, R, Ac, r_{type}, \rightarrow, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act}) \) where

(i) \( U \) is a finite set of actors/users ranging over by \( u \),

(ii) \( R \) is the finite set of artifacts/resources ranging over by \( r \),

(iii) \( Ac \) is a finite set of activities,

(iv) \( r_{type}: R \rightarrow \mathbb{R} \) is the artifact type function mapping each artifact to an artifact type

(v) \( \rightarrow: R \rightarrow R \) is parent artifact function, which is a partial function mapping artifacts to their parent artifact if defined,

(vi) \( \rightarrow_{post}: U \rightarrow \mathcal{P}_{disj}(R) \) is a partial function mapping actors to mutually disjoint subsets of artifacts,

(vii) \( \rightarrow_{share} \subseteq U \times R \) is a relation mapping users to artifacts,

(viii) \( \rightarrow_{like} \subseteq U \times R \) is a relation mapping users to the artifacts indicating the artifacts liked by the users,

(ix) \( \rightarrow_{tag} \subseteq U \times R \times \mathcal{P}(U \cup K_e) \) is a tagging relation mapping artifacts to power sets of actors and keywords indicating tagging of actors and keywords in the artifacts, where \( K_e \) is set of keywords defined in Def. ,

(x) \( \rightarrow_{act} \subseteq R \times Ac \) is a relation mapping artifacts to activities.
Definition. In Social Data $S = (G, T)$, we define Social Text as $T = (To, Ke, Pr, Se, \rightarrow_{\text{topic}}, \rightarrow_{\text{key}}, \rightarrow_{\text{pro}}, \rightarrow_{\text{sen}})$ where

(i) $To, Ke, Pr, Se$ are finite sets of topics, keywords, pronouns and sentiments respectively,

(ii) $\rightarrow_{\text{topic}} \subseteq R \times To$ is a relation defining mapping between artifacts and topics,

(iii) $\rightarrow_{\text{key}} \subseteq R \times Ke$ is a relation mapping artifacts to keywords,

(iv) $\rightarrow_{\text{pro}} \subseteq R \times Pr$ is a relation mapping artifacts to pronouns,

(v) $\rightarrow_{\text{sen}} \subseteq R \times Se$ is a relation mapping artifacts to sentiments.
Formal Model - Actions

**Definition.** In Social Data $S = (G, T)$ with $G = (U, R, A_c, r_{type}, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act})$, we define a post operation of posting a new artifact $r$ ($r \notin R$) by an user $u$ as $S \oplus_p (u, r) = (G', T)$ where $G' = (U', R', A_c, r_{type}, \rightarrow_{post}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act})$,

(i) $U' = U \cup \{u\}$

(ii) $R' = R \cup \{r\}$

(iii) $\rightarrow_{post}' = \begin{cases} \rightarrow_{post} (u) \cup \{r\} & \text{if } \rightarrow_{post} (u) \text{ defined} \\ \rightarrow_{post} \cup \{u, \{r\}\} & \text{otherwise} \end{cases}$

**Definition.** Let Social Data be $S = (G, T)$ with $G = (U, R, A_c, r_{type}, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act})$, the comment operation on an artifact $r_p$ ($r_p \in R$) by an user $u$ for a new artifact $r$ is formally defined as $S \oplus_c (u, r, r_p) = (G', T)$ where $G' = (U', R', A_c, r_{type}, \rightarrow_{post}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act})$,

(i) $S \oplus_p (u, r) = (G'', T)$ where $G'' = (U', R', A_c, r_{type}, \rightarrow_{post}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act})$,

(ii) $\rightarrow' = \rightarrow \cup \{r, r_p\}$

**Definition.** In a Social Data $S = (G, T)$ with Graph $G = (U, R, A_c, r_{type}, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act})$, we define the like operation by an user $u$ on an artifact $r$ as $S \oplus_l (u, r) = (G', T)$ where $G' = (U \cup \{u\}, R, A_c, r_{type}, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like} \cup \{(u, r)\}, \rightarrow_{tag}, \rightarrow_{act})$.

Similarly, we also define the unlike operation on $S = (G, T)$ with Graph $G = (U, R, A_c, r_{type}, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act})$, as $S \ominus_l (u, r) = (G', T)$ where $G' = (U, R, A_c, r_{type}, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like} \setminus \{(u, r)\}, \rightarrow_{tag}, \rightarrow_{act})$. 
Facebook Post - Example

\[ S = (G, T) \text{ where} \]
\[ G = (U, R, Ac, r_{type}, >, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act}) \]
\[ T = (To, Ke, Pr, Se, \rightarrow_{topic}, \rightarrow_{key}, \rightarrow_{pro}, \rightarrow_{sen}) \]
\[ Ac = \{promotion\}, \]
\[ To = \{summer collection\}, \]
\[ Ke = \{H&M, Summer\} \]
\[ Pr = \{We, I\}, Se = \{+, 0, -\}, \]
\[ U = \{u_0, u_1, u_3\}, \rightarrow_{act} = \{(r_1, promotion)\} \]

**post action by \( u_0 \)**
\[ S \oplus_p (u_0, r_1) = S_1 = (G_1, T) \text{ where} \]
\[ G_1 = (U_1, R_1, Ac, r_{type}, >, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act}) \text{ where} \]
\[ U_1 = U \cup \{u_0\}, R_1 = R \cup \{r_1\} \text{ and} \]
\[ \rightarrow_{post_1} = \rightarrow_{post} U \{(u_0, \{r_1\}\}\} \]

**like action by \( u_2 \)**
\[ S_1 \oplus_l (u_2, r_1) = S_2 = (G_2, T) \text{ where} \]
\[ G_2 = (U_2, R_1, Ac, r_{type}, >, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act}) \text{ with the following values} \]
\[ U_2 = U_1 \cup \{u_2\}, \text{ and } \rightarrow_{like} = \rightarrow_{like} U \{(u_2, r_1)\} \]

**comment action by \( u_3 \)**
\[ S_2 \oplus_c (u_3, r_2, r_1) = S_3 = (G_3, T) \text{ where} \]
\[ G_3 = (U_3, R_2, r_{type}, Ac, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{act}) \text{ with the following values} \]
\[ U_3 = U_2 \cup \{u_3\}, R_2 = R_1 \cup \{r_2\}, \rightarrow_{post_2} = \rightarrow_{post} U \{(u_3, \{r_2\}\),
\[ \rightarrow_{like} = \rightarrow_{like} U \{(r_2, r_1)\}\]
H & M Facebook Data set

- H&M Swedish fast fashion retail clothes company
- 2009/01/01 to 2013/12/31
- Total entries: 12.60 Million
- 9.95 Million likes
- 112,000 posts,
- 300,000 comments
- Albums + comments & likes on
  Albums: 2.26 Million

Pie chart showing distribution:
- Likes: 79%
- Albums: 18%
- Posts: 1%
- Comments: 2%
Artifact (text) can be analysed by machine learning tools such as Google Prediction API\(^1\)

default sentiment label: positive (+) /neutral (0)/negative (-)

a sentiment score such as \{(+):82, (0):15, (-): 03\}

Only posts and comments can have sentiments

likes and shares carry forward their parent artifact’s sentiment

---

Artificial Sentiment Distribution

Distribution of post artifacts based on the post and comments' sentiments

Temporal distribution of Artifact sentiments (quarterly)

Artifact sentiment for a given time period \((t_1, t_2)\):

\[
R^{se}_{t_1-t_2} = \{ r \mid (r, se) \in \rightarrow_{sen} \land (t_1 \leq Time(r) \leq t_2) \}.
\]
Actor Sentiments - Actor Profiling

- Actors don't carry any direct sentiment
- **Actors** perform **Actions** on **Artifacts**
- Actor sentiment can be derived from their actions on artifacts
- Total: 3.8 million unique users

Set of actors belong to a sentiment label:

\[ U_{R^{se}} = \{ u \mid \exists r \in \rightarrow_{post} (u) \land r \in R^{se} \} \cup \{ u \mid \exists r \in R^{se} \land (u, r) \in (\rightarrow_{share} \cup \rightarrow_{like}) \} \]
Actor Sentiments - Actor Profiling - II

Some of the peaks correspond to real-world events such as Factory collapses in Bangladesh
- Rana Plaza incident (week 2013-17 - 2013-20) where 1129 people died
- Seven people killed in week 2013-41
H & M Sales - Artifact Sentiments

- Strong correlation between sales and +ve comments on non-H&M posts
- Strong correlation between sales and -ve posts by non-H&M
- Strong correlation between sales and -ve comments on non-H&M posts
- Strong correlation between sales and neutral posts by non-H&M
Future Work

• Formal model can abstracted further to model data from other social media channels such as Twitter

• Use of fuzzy sets and fuzzy logic to develop advanced analysis techniques

• Modeling of networks of groups and friends of users in an online social media platform

• More case studies to study consumer behaviour in case of crisis events

Questions & Comments?