International Journal of Computer Science and Business Informatics

IJCSBI.ORG

A Survey on Knowledge Analytics of Text from Social Media

Dr. J. Akilandeswari
Professor and Head,
Department of Information Technology
Sona College of Technology,
Salem, India.

K. Rajalakshm
PG Scholar, Department of Information Technology
Sona College of Technology,
Salem, India.

ABSTRACT
Actionable knowledge discovery is a closed optimization problem solving process from problem definition. It is used to extract the actionable data that are usable. Social media still contain many comments that cannot be directly acted upon. If we could automatically filter out such noise and only present actionable comments, decision making process will be easier. Automatically extracting actionable knowledge from on line social media has been attracted a growing interest from both academia and the industry. This paper gives a study in the systems and methods available text from the social media like twitter or Facebook.

Keywords

knowledge discovery, social networking, classification.

1. INTRODUCTION

Social networking becomes one of the most important parts of our daily life. It enables us to communicate with a lot of people. Social networking is created to assist in online networking. These social sites are generally communities created to support a common idea. Data mining is the process of discovering actionable information from large sets of data. Actionable knowledge discovery from user-generated content is a commodity much sought after by industry and market research. The value of user-generated content varies significantly from excellence to abuse. As the availability of such content increases, identifying high-quality content in social sites based on user contributions is very difficult. Social media sites become increasingly important. In general social media demonstrate a rich variety of information sources. In addition to the content itself, there is a large array of non-content information obtainable in these sites, such as links between items and unambiguous quality ratings from members of the community. We argue that to achieve the goal we must gain a better understanding of what actionable knowledge is, where it can be found and what kind of language structures it contains. The aim of this work is to do so by analyzing actionable knowledge in on-line social media conversation.

2. Related works

Maria Angela et al., [2] has proposed understanding Actionable knowledge in social media BBC Question time and twitter. This paper will answer the following questions: What is actionable knowledge, whether it can be measured and where can we find for gaining better understanding of actionable knowledge in twitter? There are three types of tweets: closed, re-tweet, open. Actionable tweets can found in any of these categories. Three steps are involved; 1) manually annotate the three subsets with action ability scores. 2) Test the hypotheses by performing statistical annotated data. 3) Use the W Matrix to automatically identify the language patterns in actionable data. The method used in this paper prepares two sets Seta containing actionable data and sets containing non actionable data. The two sets of data are then loaded into the W matrix.

Eugene Agichtein et al., [3] have proposed to automatically asses the quality of questions and answers provided by the user of the system. They take the test case as Yahoo! Answers. They introduce the general classification framework for combine the substantiation from different sources of information, which can be adjusted automatically for a given social media type and quality definition. Sub problem of quality evaluation is an essential module for performing more advanced information retrieval tasks on the question/answering. The interactions of users are organized around questions like 1) asking a question 2) answering a question 3) selecting best answers 4) voting on an answer.

Models:

- Intrinsic content quality: The content quality of each item. This is mostly used text related.
    - Punctuation and typos
    - Syntactic and semantic complexity
    - Grammatically
- Usage statistics: Clicks on the item.

Modeling content quality in community Question/Answering:

Application-specific user relationships:

The dataset, viewed as a graph, contains multiple types of nodes and multiple types of interactions

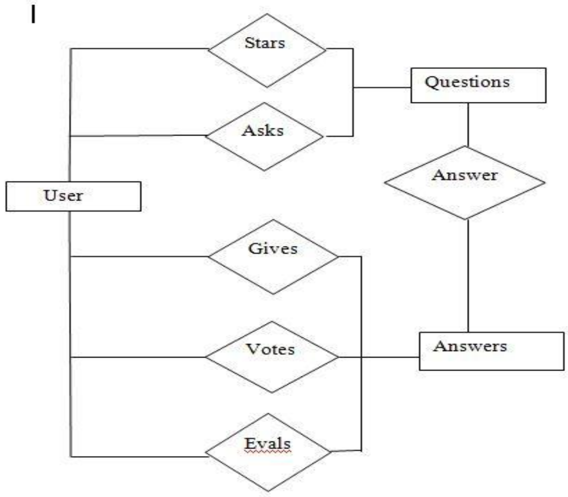

Fig 2.1 Partial Entity-Relationship Diagram for Answers

The relationships between questions, user asking and answering questions, and answers can be captured by a tripartite graph outlined in the figure where an edge represents an explicit relationship between the different node types. Since a user is not allowed to answer his/her own questions.

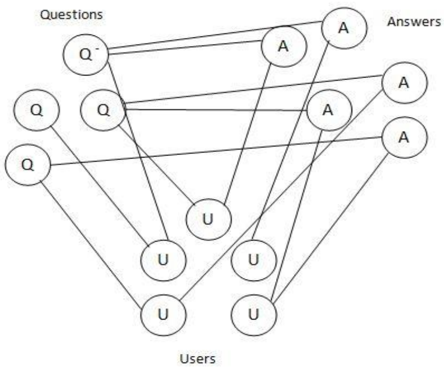

Fig 2.2 Interaction of user-questions-answers modeled as a Tri-partiate Graph.

The types of features on the question sub tree:

Q represents features from the question being answered.

QU represents features from the asker of the question being answered.

QA represents features from the other answer to the same question.

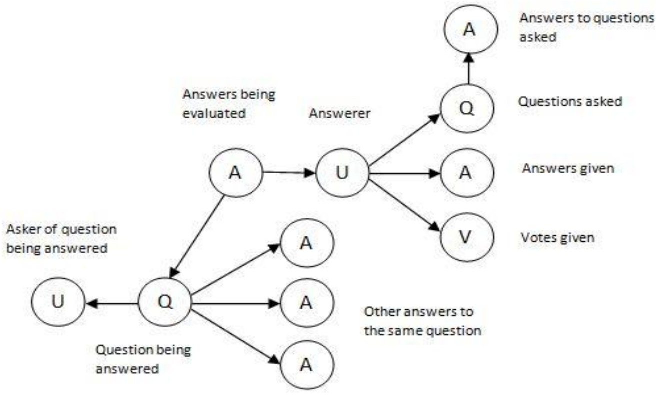

Fig 2.3 Types of features available for inferring the quality of question.

The types of features on the user sub tree:

UA represents features from the answers of the user

UQ represents features from the question of the user

UV represents features from the votes of the user

UQA represents features from answers user received to the user's question.

U represents other user based features.

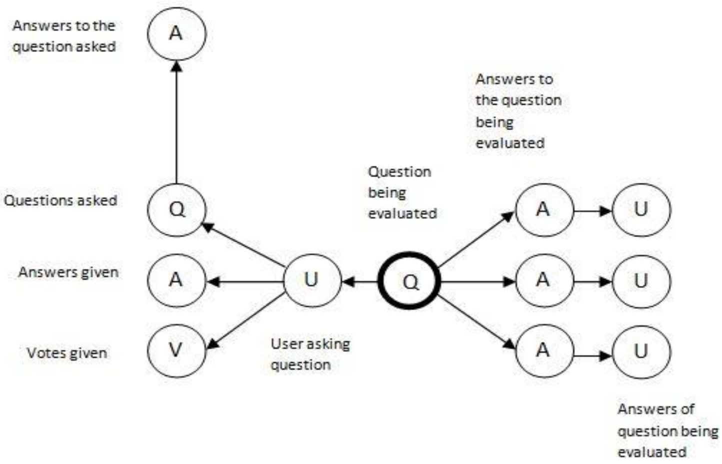

Fig 2. 4 Types features available for inferring the quality of a question

A represents feature directly from the answer received.

AU represents features from the answers from the question being answered.

Akshay et al., [4] has proposed about why we use the twitter and understanding of microblogging usage and communities. The growth of the twitter is increased because of microblogging. It is a new form of communication in which users can describe their current status in short posts distributed by instant messages. One of the microblogging platform is twitter. This tool provides a light-weight easy form of communication that enables user to broadcast and share the information about their activities. Compared to regular blogging, microblogging fulfills a need for an even faster mode communication. Frequency update is one of the differences between the regular blogging and microblogging. On an average, a profile blogger may update her blog once every few days, but microblogger may post several updates in a single day. Microblogging allow only 140 characters. This paper proposes a two level of framework for user intention detection. 1) HITS algorithm to find hub and authority network. 2) Identification of communities within friendship wise relationships by only considering the bidirectional links where two users regard each other as friends. The main user intention in twitter is Daily chatter, conversation, sharing information, reporting the news.

Swapna Gottipati and Jing Jiang [5] proposed extraction of entity-actions from users' commentary. Opinion mining process focuses on extraction of sentiments on social, products, political and economical issues. In many cases, users not only express their sentiments but also share their requests ideas, and suggestions through comments. This paper defines a new problem that is extracting entity-actionable knowledge from the user's explanation.

Example:

- Government must lift diplomatic immunity of the ambassador.
- Government must inform the Romanian government of what happened immediately.
- SG government wants to cooperate closely with Romania in persecuting this case.
- Hope the government helps the victims by at least paying the legal fees.
- I believe that government will help the victims for legal expenses.

The above comments are in response to the news about a car accident.

- First, all sentences consist of an action and the corresponding entity who should take the action.
- Second, users tend to express the actions in various sentence structures and hence extracting entities and actions is desired and challenging as well.
- Third, we observe that entities in all the above sentences refer to the same entity but expressed in various forms.
- Finally, similar actions are expressed differently which drives the need for normalizing the action.

Entity action extraction:

There are three main properties of actionable comments

- The entities of actionable pairs are mostly nouns or pronouns.
- The entities display the positional properties with respect to the keyword.
- Entities grammatically related to the action. Eg) verb in the phrase is related to the subject which is an entity of the actionable comment.

Table 3.1 Sample output of actionable comments extraction and normalization task

Entity	**Action**
Govt	Lift diplomatic immunity of the ambassador and get him to face.
Govt	Inform the Romanian government of what happened immediately.
Govt	Cooperate closely with Romania in persecuting.
Govt	Helps victims by at least paying the legal fees.

Zengyou He et al., [6] given the survey of Data mining for actionable knowledge Data mining consists of series of steps are:

- Data selection
- Data cleaning
- Data transformation

Actionable refers to the mined patterns suggest profitable actions to the decision maker. The user can do something to bring direct benefits. The benefits are:

- Increase in profits
- Reduction in cost
- Improvement in efficiency.

There are two frameworks used for mining the actionable knowledge. Loosely coupled framework :

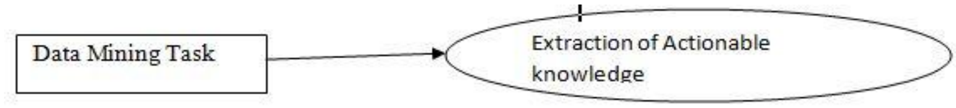

Fig 3.7 The general procedure to go from data mining task to actionable knowledge in loosely coupled framework.

Advantages of loosely coupled framework: Flexibility, Independencies on application.

Tightly coupled framework is better than the loosely coupled framework in finding actionable knowledge to maximize profits

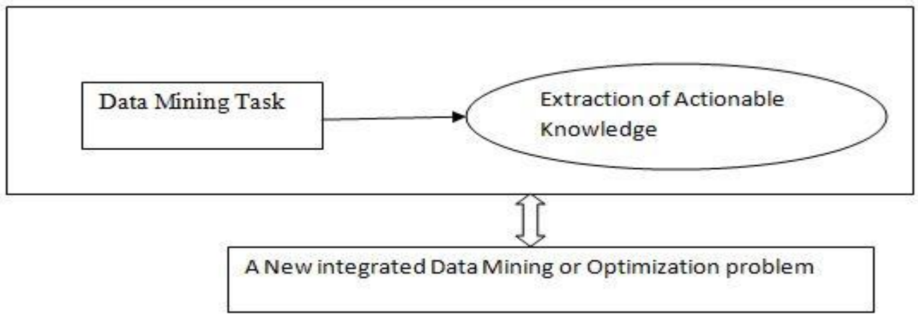

Fig 3.8. The general procedure to go from data mining task to actionable knowledge in tightly coupled framework.

Disadvantages of tightly coupled framework: It is strongly dependent on the application domain, The new formulated problem is usually very complex.

Killan Thiel et al., [7] has proposed the predictive analytic techniques and text mining the Slashdot data. Predictive analytic technique used on social media enables the user to start generate new fact –based approaching on the social media data. Text mining has been used to do sentiment analysis on social media data. Sentiment analysis takes the written content and translated it into different contexts, such as positive and negative. Sentiment analysis depends on an suitable subjectivity lexicon that understands the virtual positive, neutral or negative perspective of word expression.

Example: I find PRODUCTX to be good and useful, but it is too expensive

The term (and therefore the PRODUCTX) is rate as positive, since there are two positive expressions "good" and "useful" – and one negative word "expensive". In addition, one of the positive word is improved with the word "very" while the negative word is place into perspective by the qualifier "a bit". The more highly developed lexica, the more detailed analysis and the findings can be. Sentiment analysis using text mining can be very powerful and is a well-establish, stand-alone predictive analytic technique.

In a first step, identify negative and positive users, that which paper is to establish whether the known (not anonymous) users express predominantly positive or negative opinions, feelings, attitudes, or sentiments in their comments. The polarity is used to categorize sentiment lexicon containing texts (clues). The polarity of a text specifies whether the word seems to evoke positive or negative. probable polarity values are: positive, negative, both, and neutral. KNIME is used to choose the lexicon.

Danah boyd et al., [8] examines the practice of retweeting. The goal of this paper is to describe and map out the various convention of retweeting to provide the framework for examining retweeting practices. Retweeting practices are:

How people retweet: Twitter gives the option to retweet through a simple click, or you can type "RT" followed by a space and the "@username" of the original author. Retweeting is used to share the valuable content, build the relationship, Retweeting information regularly is to enhance our reputation and establish our authority.

Why people retweet: To increase or spread tweets to new audience.

- To entertain a particular audience.
- To comment someone's tweet by retweeting.
- To make one's presence of listener visible.
- To publicly agree with somebody.
- To certify other thoughts.
- To identify less popular people or less visible content.

What people retweet: retweet for others and retweet for social action.

Pritam Gundecha and Huan Liu [9] had given brief introduction about mining the social media. Mining social media has its potential. To extract actionable pattern that can be beneficial for users, business, and consumers. Social media data are noisy, vast, unstructured, and thus novel challenges arise. Data mining of social media can expand researchers' capability of under-standing new phenomena due to the use of social media and improve business intelligence to provide better services and develop innovative opportunities. For example, data mining technique can help identify the influential people in the vast blogosphere, detect hidden groups in a social networking site.

Issues in mining the social media are:

Community analysis: A community is formed by individuals such that those groups interact with each other more frequently than with those outside the group. Communities classified into two groups implicit group’s explicit groups. Explicit groups are formed by user subscriptions, and implicit groups are formed by user interaction. Community analysis faced with issues such as community detection, formation and evaluation. The main challenges in community detection are 1) the definition of a community can be subjective. 2) the lack of ground truth makes community evaluation difficult.

Sentiment analysis and opinion mining: Sentiment analysis and opinion mining is used to automatically extract opinions expressed in the user-generated content. Opinion mining and sentiment analysis tools allow businesses to understand brand perception, product sentiments, new product perception, and reputation management. Sentiment analysis is hard because languages used to create contents are ambiguous. the steps of sentiment analysis are:

Finding the relevant sections,

Finding relevant documents,

Finding overall sentiment,

Quantifying the sentiment,

Aggregating all sentiments to form an overview.

They proposed the tool twitter tracker is a twitter based analytic and visualization tool. The focus of the tool is to help HADR relief organization to acquire situation awareness during disasters and emergencies to aid disaster relief efforts. New social media platforms such as twitter microblogs, demonstrate their value and capability to provide the information that is not attainable in traditional media. Twitter tracker is designed to track, analyze, and monitor tweets.

Yanchang Zhao et al., [10] have proposed a Combined Pattern Mining method. Association mining produces great collections of association rules that are hard to understand and put into action.

Combined Association Rule: Assume that there are k datasets Di (i = 1…..k). Assume Ii to be the set of all items in datasets Di and ∀i ≠ j, Ii ∩ Ij = ∅. A combined association rule R is in the form of A1 ∧ A2 ∧ ... ∧ Ak → T, where Ai ⊆ Ii (i = 1...k) is an itemset in dataset Di, T ≠ ∅ is a target item or class and ∃i, j, i ≠ j, Ai ≠ ∅, Aj ≠ ∅. For example, A1 is a demographic itemset, A2 is a transactional itemset on marketing campaign, A3 is a an itemset from a third-party dataset, and T can be the loyalty level of a customer. The combined association rules are then further organized into rule pairs by putting similar but contrasting rules together.

Combined Rule Pair: Assume that R1 and R2 are two combined rules and that their left sides can be split into two parts, U and V , where U and V are respectively itemsets from IU and IV (I = {Ii}, IU ⊂ I, IV ⊂ I, IU ≠∅, IV ≠∅ and IU ∩ IV = ∅). If R1 and R2 share a same U but have different V and different right sides, then they build a combined rule pair P as

R1 : U ∧ V1 → T1

R2 : U ∧ V2 → T2

where U ≠∅, V1 _= ∅, V2 ≠∅, T1 ≠ ∅, T2 ≠∅, U ∩V1 = ∅, U∩V2 = ∅, V1∩V2 = ∅ and T1 ∩ T2 = ∅.

A combined rule pair is composed of two contrasting rules, which suggests that for customers with same characteristics U, different policies/campaigns, V1 and V2, can result in different outcomes T1 and T2. Based on a combined rule pair, related combined rules can be organized into a cluster to supplement more information to the rule pair.

Mario Cataldi et al., [11] has recognized primary role of Twitter and propose a novel topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the community. First they select the emerging terms from that select the emerging topics. They propose two techniques that are a supervised and an unsupervised technique to select a limited set of relevant terms that emerge in the considered time interval. In supervised selection they introduce critical drop value that allows the user to decide when a term is emergent. In particular, the critical drop is defined as

$$drop^t = \delta . \frac{\sum k \epsilon k^{t(energy\ k^t}}{|k^t|}$$

where δ ≥ 1. It permits to set the critical drop by also taking into account the average energy value. Therefore, we define the set of emerging keyword EKt as,

$\forall k \in k^t, k \in Ek^t \leftrightarrow energyk^t > drop^t$

The second approach considers a completely automatic model that does not involve any user interaction. Unsupervised ranking model dynamically sets the *critical drop.* This cut-off is adaptively computed as follows.

1. First ranks the keywords in descending order of energy

Value calculated.

2. Computes the *maximum drop* in match and identifies the corresponding drop point.

3. Computes the *average drop* (between consecutive entities) for all those keywords that are ranked before the identified maximum drop point. The first drop which is higher than the computed average drop is called the *critical drop*.

We can define the topic as a minimal set of terms semantically related to an emerging keyword. Thus, in order to retrieve the emerging topics, we consider the entire set of tweets generated by the users within the time frame. for example, the keyword "*victory*" in a given set of tweets: this term alone does not permit to express the related topic. In fact, considering as a time frame November 2008, the related topic can be easily defined by the association with other keywords (among the most used) as "elections", "Usa", "Obama" and "McCain". By using correlation vector we can identify the topics related to the emerging terms retrieved during the considered time interval. Topic graph and topic detection and ranking algorithm is used to find emerging topic.

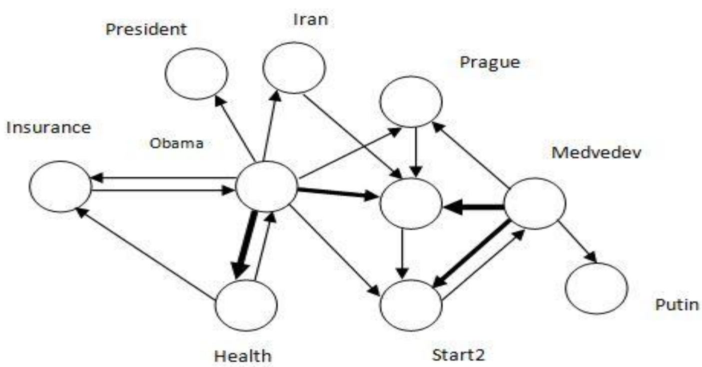

Fig 3.10 A Topic graph with two Strongly Connected

Components representing two different emerging topics: labels in bold represent emerging keywords while the thickness of an edge represents the semantical relationship between the considered keywords.

Pirooz Shamsinejad and Mohamad sararee [12] has proposed a Bayesian network approach. It used for causal action rule mining. Action rule is a new method in this research area. It suggest some actions to user to get a profit in his/her domain. classification rules, Decision trees and association rules are already used for action rule mining. But these are all not used for the causal relationship. Bayesian network shows the causal relationship between variables of interest for extracting action rule. Causal relationship is one variable causes a change in another variable.

Action rule Discovery using Bayesian network: This paper proposed action discovery method based on Bayesian network. There are two phases 1) Modeling phase: It takes data about the instances as input and then creates a Bayesian Network (BN) for modeling causality relationship between attributes of instances in database. 2) Discovering phase method: It takes each time an instance and generates the highest profitable actions for that case.

In modeling phase there are three steps:

1. Specify the set of relevant attributes and their values. It means defining the domain of problem.

2. Construct the structure of BN by connecting each pair of attributes for which there is a cause and effect relationship between them. The resulting structure is a Directed Acyclic Graph (DAG) whose nodes are attributes and each causal relationship is shown by an edge starting from cause node pointing to effect node

3. Learning the parameter which means finding the values of conditional probabilities for each attribute. It is the quantitative part of the learning process.

Discovery phase:

Step1: Find the candidate action rules for the instance.

Step2: Estimate the power of each action rule that to be Change the goal attribute

Step3: Ranking the action rules based on their power and Selecting the most profitable ones.

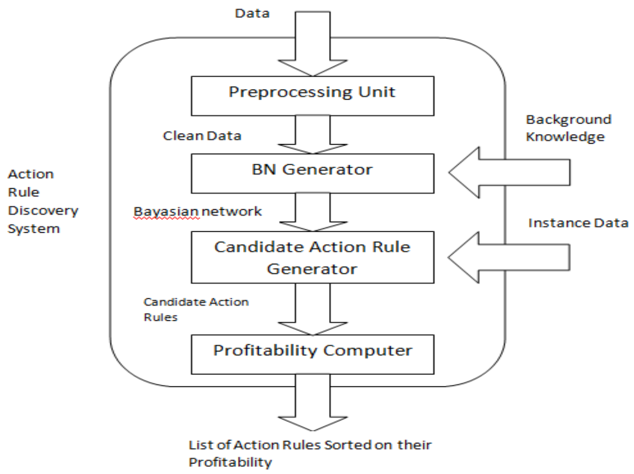

Fig 3.5 Action rule Discovery system based on Bayesian networks

Courteney Honeycutt and Susan C.Herring [13] has given the detail about conversation and collaboration via twitter. The microblogging twitter is in the process of being appropriated for conversational interaction and is starting to be used for collaboration. The goal of this paper is to collect all tweets posted to the public timeline. Twitter scraper is one of the tool is available for public use. It is used to collect the data. In this paper twitter scraper is used collect the conversation in twitter. All the tweets are not collected because of two reasons. 1) Twitter scraper was only able to collect up to 20 tweets per operation. 2) During period of heavy activity twitter scraper takes long time to gather data or return error messages.

Longbing Cao and Chengqi [14] propose a practical perspective, referred to as a domain-driven in-depth pattern discovery (DDID-PD). It corresponds a domain driven view of of discovering knowledge satisfying real business needs. It includes constraint mining, human cooperated mining, in-depth mining, and loop closed mining.

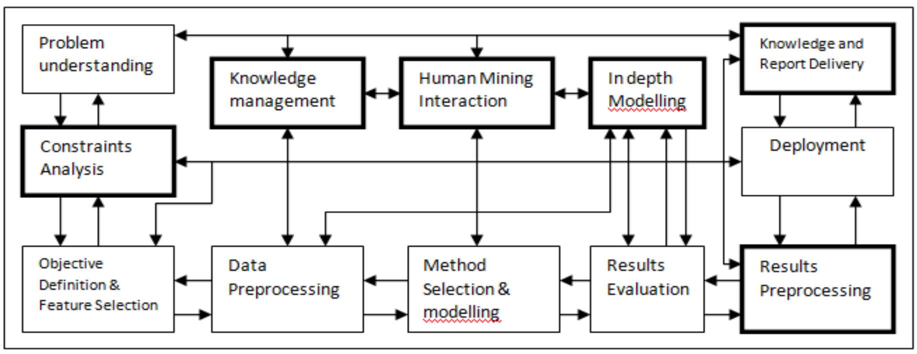

Fig 3.9 DDID-PD process model

Constraint mining: Several types of constraints are listed, which play the important roles in a process efficiently discovering knowledge actionable to business.

In –depth mining: In in-depth mining, attentions should be paid to business requirements, domain knowledge, objectives, and qualitative intellect of domain expert for their impact on mining deep pattern. This might be done through select and adding up business features, consider domain and background data in modeling, supporting interact with fine tuning parameters domain experts, and data set by domain experts, parameters, optimizing models and adding factors into technical interesting measures or building business measures, improving those result evaluation mechanism through embedding human involvement and domain knowledge, etc.

Human cooperated mining: In DDID-PD the role of human could be personified in the full period of data mining from business and data accepting problem definition, data assimilation and sampling, hypothesis proposal, feature selection, business modeling and learning to the evaluation and resulting outcomes.

Loop-closed mining: it encloses iterative feedback to varying stages such as sampling, hypothesis, feature selection, modeling, evaluation and interpretation in a human-involved manner.

Janaina et al., [15] has proposed dengue surveillance based on computational model of spatio-temporal locality of twitter. how much, where and when dengue incidence happened, but also an additional dimension enabled by social media, which is how the population faces the epidemics. Then introduce an active surveillance framework that analyzes how social media reacts epidemics based on four dimensions: volume, time, location, and public perception.

Aron Culotta [16] has proposed detecting influenza epidemics by analyzing twitter messages. He analyzed messages posted on the microblogging site twitter.com. He proposed several methods to identify influenza related messages and compare a number of regression models to correlate these messages with CDC statistics.

Modeling Influenza rates:

He considered P be the true proportion of the population exhibiting ILI symptoms. And W= {w1…..wk}be a set of k keywords. He considered D as a document collection and Dw as the set of documents in D that contain at least one keyword in W. He defined Q(W,D) = $\frac{|Dw|}{|D|}$ was the fraction of documents D that match W.

Regression Models:

Simple Linear Regression:

He considered a simple linear model between the log-odds of P and Q(W,D):

$$\text{Logit(P)} = \beta_1 \text{logit(Q(W,D))} + \beta_2 + \varepsilon$$

With coefficient β_1 , β_2 , error term ε and logit function logit(X) = $\ln(\frac{X}{1-X})$.

Multiple Linear Regression :

W contained more than one keyword, it is natural to considered expanding Simple Linear Regression to include separate parameters for each element of W. The results of multiple Regression model:

Logit(P) = β_1 logit (Q({w1},D)) +………..+β_k logit(Q{wk}, D)) + β_{k+1} + ε where w_iε W.

Keyword Selection:

Keyword was selected by using correlation coefficient and Residual Sum of Squares.

Key word Generation:

He proposed two methods to generate the keywords.

Hand chosen keywords:

He considered a simple set of four keywords consisting of {flu , cough , sore throat , headache}.

Most frequent keywords:

To expand this candidate set, he searched for all documents containing any of the hand-chosen keywords. Then he found 50,000 most frequently occurred words in the resulting set.

4. CONCLUSION

The paper presented techniques which analyze actionable knowledge in on-line social media conversation. We clarified the notion of actionability by considering the actionability of those expressions (or tweets) that contain a request or a suggestion that can be acted upon. We have identified actionability in dengue related tweets. For example, let's examine the following two tweets: A. protect your children from dengue. B. dengue is a disease. Both tweets contain an opinion about the same topic (e.g. society/social innovation). However, tweet A suggests a clear action. We argue that tweet A says something than tweet B, in that A contains actionable knowledge and B does not. We believe that it is important to understand not only *how people feel* about a topic but also *what actions* they would like to take. To infer actions from the tweets, soft computing techniques are to be employed with the training keyword set.

5. REFERENCES

[1] Zhao, D. and Rosson, M.B, 2009. "How and why people Twitter: the role that micro blogging plays in informal communication at work", In Proceedings of the Int. Conference on Supporting Group Work (GROUP '09).

[2] Maria Angela Ferrario, Will Simm "Understanding Actionable Knowledge in Social Media", Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[3] Agichtein, E., Carlos, C., Donato, D., Gionis, A. and Mishne, G, "Finding high quality content in social media", In Proceedings of The International Conference on Web Search and Web Data Mining (WSDM '08). ACM, New York, NY, USA, 2008.

[4] Java, A., Xiaodan, S., Finin, T. and Tseng, B, "Why we Twitter: Understanding Microblogging Usage and Communities", In Proceedings of the 9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis (WebKDD/SNA KDD '07). ACM, New York, NY, USA, 56 65, 2007.

[5] Swapna Got t ipat i, J ing J iang School of Information Systems, Singapore Management University, Singapore " Extracting and Normalizing Entity-Actions from Users' Comments" Proceedings of COLING : Posters, pages 421–430, 2012.

[6] Zengyou He,Xiafei Xu,Shengchun Deng, "Data mining for Actionable Knowledge: A Survey" , research and development program of china and the IBM SUR Research Fund, 2003.

[7] Killian Thiel Killian, Tobias Kötter , Dr. Michael Berthold Michael, Dr. Rosaria Silipo Rosaria,Phil Winters, "Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining" by KNIME.com AG all rights reserved 2012.

[8] Boyd, Danah, Scott Golder, and Gilad Lotan,. "Tweet, Tweet, Retweet:Conversational Aspects of Retweeting on Twitter." HICSS-43. IEEE: Kauai, HI, January 6, 2010.

[9] Pritam Gundecha, Huan Liu Arizona State University, Tempe, Arizona 85287 "Mining Social Media: A Brief Introduction" ,Tutorials in operation research informs 2012.

[10] Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Chengqi Zhang, and hans Bohlscheid, " Combined Pattern Mining: From Learned Rules to actionable knowledge", AI, LNAI 5360, pp.303-403, 2008.

[11] Cataldi, M., Di Caro, L. and Schifanella, C, "Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation", In Proceedings of the Tenth International Workshop on Multimedia Data Mining (MDMKDD '10). ACM, New York, NY, USA, Art. 4, 2010.

[12] Pirooz Shamsinejad and Mohamad Saraee," A Bayesian Network Approach for Causal Action Rule Mining", International Journal of Machine Learning and Computing, Vol. 1, No. 5, December 2011.

[13] Honeycutt, C. and Herring, S.C. "Beyond Microblogging: Conversation and Collaboration via Twitter", In Proceedings of 42nd Hawaii International Conference on System Sciences. (2009) pp. 1 10, 2010.

[14] Cao, L. and Zhang, C, " Domain Driven Actionable Knowledge Discovery in the Real World", Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 821 830, 2010.

[15] Janaina Gomide , Adriano veloso , Wagner Meira Jr , Virgilio Almeida , Fabrico Benevenuto , "Dengue Surveillance based on a computational model of spatio-temporal locality of twitter"

[16] Arone Culotta , "Towards detecting influenza epidemics by analyzing Twitter messages" 1st workshop on social media analytics , July 25, 2010.

This paper may be cited as:

Akilandeswari, J. and Rajalakshm, K., 2014. A Survey on Knowledge Analytics of Text from Social media. *International Journal of Computer Science and Business Informatics, Vol. 13, No. 1, pp. 45-61*.