Evolving social data mining and affective analysis methodologies, framework and applications

Athena Vakali
Informatics Department, Aristotle University
Thessaloniki, Greece
2310998415, 54124
avakali@csd.auth.gr

ABSTRACT

Social networks drive todays opinions and content diffusion. Large scale, distributed and unpredictable social data streams are produced and such evolving data production offers the ground for the data mining and analysis tasks. Such social data streams embed human reactions and inter-relationships and affective and emotional analysis has become rather important in todays applications. This work highlights the major data structures and methodologies used in evolving social data mining and proceeds to the relevant affective analysis techniques. A particular framework is outlined along with indicative applications which employ evolving social data analysis with emphasis on the seminal criteria of topic, location and time. Such mining and analysis overview is beneficial for various scientific and enterpreneural audiences and communities in the social networking area.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *distributed systems*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*; I.7.5 [**Document and Text Processing**]: Documents Capture – *Document analysis*

General Terms

Design, Algorithms, Performance, Experimentation, Human Factors.

Keywords

Social networking, microblogging data analysis, evolving social data mining, social affective analysis.

1. INTRODUCTION

Web has radically evolved from a static system of interlinked hypertext documents into an interactive collaborative platform for content generation, circulation and diffusion. In nowdays Web 2.0 era, users have eagerly become active content generators and regulators, and they regulate content via a variety of web applications such as: blogs, wikis, social tagging or bookmarking applications, and social networking sites. Social networks have drastically increased online communication and human interactions, since millions of users share opinions on a variety of topics. Such activities embed both objective and subjective criteria and certainly human reactions govern social media diffusion.

Massive participation in Web 2.0 applications renders the Web a source of huge sizes of social data, with the term describing the online associations between different types of actors, such as: users, resources and metadata, which are either explicit or implicit. This holds since for example, users can be explicitly connected with friendship bonds in social networking applications, but at the same time there may also exist relations implied by users' common activities, such as commenting or posting on common online resources. The relevance and interrelationships of social media actors emerge in dependency to the orientation and functionality of the social networking applications themselves, but at it is important to reveal hidden relationships, trends and phenomena via appropriate techniques and methodologies. Mining techniques, appropriately tailored for evolving social data streams, are certainly crucial. Moreover, since social media interactions are driven by human decisions emotional awareness has also an important role to play in any mining and analysis process.

Sentiment analysis in social media has gained considerable ground lately since it captures human behavioural patterns and highly contributes in understanding social media responsiveness and users reactions. Positive, negative and neural opinions are now declared, triggered and visualized in most of the current social media applications. Expressing such opinions and senses is important not only from the scientific analysis side but also from the side of markets and stakeholders since they can accordingly suggest, and apply policies and services according to the social crowd opinions.

This work addresses the challenges inherent on social networking data streams in a dual pathway: the evolving data mining and the affective analysis pathways. The focus of the work is inspired by the principle that social data streams can be analyzed to uncover hidden relations, trends and phenomena, expected to lead to often non-obvious relations and conclusions. This type of analysis requires specialized data mining and analysis methods tailored to handle: (i) the massive data sizes, (ii) the complex structure of social data's associations, (iii) the frequent data updates, and (iv) the emotions that drive social users decisions and actions. The analysis of Web 2.0 social data is a challenge for data mining which concerns both computer scientists as well as social web applications' owners and developers and market stakeholders.

The rest of the paper is organized as follows: Section 2 highlights the most popular data structures used in evolving social data representation whereas Section 3 summarizes the key mining and

IDEAS12 2012, August 8-10, Prague [Czech Republic]
Editors: Bipin C. Desai, Jaroslav Pokorny, Jorge Bernardino

affective analysis methodologies as the ones are heavily used in evolving social media analysis. Section 4 sets a particular 3-tier framework which can flexibly drive social evolving analysis applications development and two indicative applications are presented. These applications refer to trend detection and crowd emotion capturing and relevant conclusions and future challenges are highlighted in closing the work at Section 5.

2. Evolving social data structures

The graph-like network model has been considered as the obvious choice in representing social networking and Web data. In this model social data are interconnected through associations forming a network or graph G(V, E), where V is the set of nodes and E is the set of edges. The nodes represent entities/objects and edges represent relations among them [6], [8]. Different types of nodes and edges such as weighted or unweighted, directed or undirected can be exploited in dependence to the social media data at hand. On the basis of the graph model several variations have been introduced to capture social data evolution and inter-dependencies.

Figure 1. A hypergraph model for a social tagging application

Definition 1 (Hypergraph and folksonomy): A hypergraph is a generalization of a graph where an edge connects more than one nodes. A folksonomy is a lightweight knowledge representation emerging from the use of a shared vocabulary to characterize resources and it forms a tripartite hypergraph [9, 13].

Figure 1 summarizes such a folksonomy graph structure for a social tagging application which involves a number of users (n) who assign tags (totally m tags assigned) on a set of r particular resources (such as images, text etc).

Based on such hypergraph models projection processes can result in bipartite and/or unipartite graphs for simplicity in processing and analysis [1]. For example, we may result in a tag to tag graph network where two tags are connected if they have been assigned to the same resource.

Definition 2 (Adjacency and similarity matrix): to facilitate data analysis process a graph's structure G(V, E), can be encoded in an adjacency matrix A with values A[i,j] = 1, in case an edge (i,j) exist in E and G is unweighted or a similarity matrix M with values M[i,j] = w(i,j), in case an edge (i,j) exist in E and G is a weighted graph with edge weights w.

These structures have been adequate for Web static activities but it is true that evolving data demand evolving structures, and thus graph structure should change and be updated over time. Figure 2 depicts the different levels of granularities which specify data structures evolution with respect to the timeline of social data evolution and where graph snapshots, form segments to end up on a general graph stream in the upper time hierarchy level [7].

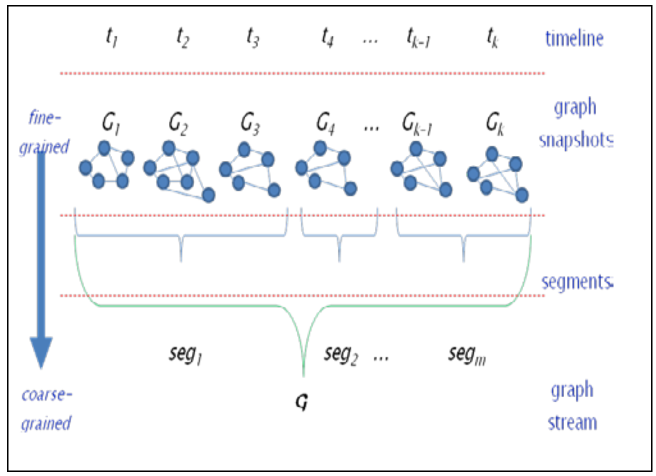

Figure 2. Graph structures over different timeline granularities

Assuming discrete time-steps, such as time-steps t and t+1 either only one update operation can take place (single update) or multiple update operations can take place (batch updates). These granularities specify distinct data structures to be used and up to now popular such structures are the tensors which represent the generalization of matrices to more than two 2 dimensions) [18] or "multi-graphs" with edges encoding relations as well as temporal information [24].

More specifically and based on the time granularity levels highlighted in Figure 2, the up to now most popular evolving data structures and models are as follows :

- *graph snapshots level :* adjacency and similarity matrices; folksonomies;
- *segments level :* adjacency and similarity matrices; tensors; pre-processing techniques to identify graph segments consisting of similar snapshots and compute a smooth graph approximation for each segment [23];
- *graph stream level :* multi-graphs; tensors.

The evolution of graph structures over time and among different time steps involves several tasks which deal with the graphs' : growth (upon arrival of new nodes), shrinkage (on departure of existing nodes), densification (on creation of new edges), Sparsification (upon deletion of existing edges) and weight updates (when weight increases or decreases for existing edges) [20].

3. Mining and affective analysis methodologies

The availability of massive sizes of data gave new impetus to data mining since mining over social web data can act as a barometer of the users' opinions and often non-obvious results may emerge. Collaboration and contribution of many individuals leads to the formation of collective intelligence and it is widely recognized that wisdom of the crowd is now considered as an accurate, unbiased source of information. In this context, we place emphasis on the methodologies dealing with evolving social data mining and affective analysis.

Various mining techniques can be used in evolving social media such as community detection, clustering, statistical analysis, classification, and association rules mining. Here we place emphasis on the clustering approaches since they have progressed in the social networking efforts and they have dominated social media analysis and processing [7], [10], [15], [16].

3.1 Evolving social data clustering methodologies

Several methodologies are well suited for clustering such evolving social data and already several approaches have been followed. These approaches are characterized by the data structures used and the model these follow, by the methodology followed to capture social data evolution and certainly by the scope and outcome of such a methodology.

An overview of representative evolving social data clustering approaches is highlighted in [7] and its brief summary is given in Table 1.

Table 1. Evolving social data clustering summary [7]

Structure/model	Social data evolution capturing	Methodology/outcome /scope
directed weighted graph snapshots for a given theme; interaction matrices aggregating interactions among current-older posts under an exponential decay scheme	events in communities structure: birth, death, propagation, growth, contraction, split, merge	users' mutual awareness drives community formation (e.g. bloggers commenting on each other's blogs); models mutual awareness expansion by a random walk process to detect communities in each snapshot; communities as vectors in an inter-blogger interaction space and correlation computed via histogram intersection to detect evolution [11].
undirected graph snapshots; affinity matrices	changes in community membership	blog communities changesbecause of concept drift; proposed two frameworks that evaluate the clustering quality on each snapshot by: partitioning to historic data, and comparing with historic partitioning schemes [3].
undirected unweighted graph and change stream	critical time-points when communities split or merge	clustering depends on maximal clique set; each change maps to a sequence of changes of the clique graph, affecting only the cliques containing the respective nodes [5].
undirected weighted graph snapshots with edges aggregating past weights under an exponential decay scheme	changes in communities structure: birth, death, propagation, growth, contraction, split, merge; changes in number of communities	found overlapping communities based on *clique percolation*; successive communities mapping by relative nodes' overlap; communities lingering in time are small and *stationary* or large and *non-stationary* [17].
undirected weighted graph and change stream; incidence matrices (some representation of the Laplacian matrix)	detected more or less stable communities with fluctuating membership for some mutli-topic blogs	decomposes node additions /deletions in weight changes that are appended as vectors to the incidence matrix; graph partitioning by minimization of a criterion based on normalized cut; updates eigenvalues and vectors via closed-form and approximate solutions respectively, and iteratively refines them; [Ning10].

3.2 Emotion-aware social data analysis

Social networks embed human reactions and capturing emotional trends out of social networking activity has proven too important. The goal of affective analysis is to capture the wisdom of the crowds, as well as the social pulse and the trends, since information is valuable for improving social network services in a wide spectrum of applications. To support such analysis, methodologies have extended ideas in social data mining to new forms which focus on capturing sentiments polarities and intensities.

Affective evaluations over social networking content have been employed typically via sentiment analysis. Sentiment Analysis (or opinion mining) refers to the computational study of opinion, sentiment and emotion of a resource and it has been typically used in text corpuses [12] and this area has importantly grown up in social networking data streams [21].

Most of the previous research work proposed methods for the sentiment classification of product or movie reviews coming mostly from forums and blogs [19]. Moreover, there is a number of interesting efforts which have used microblogging sources (such as Twitter) as a source for sentiment analysis [15], [21]. In such efforts sentiment word frequencies in tweets, and appropriate text stream mining is used to detect users opinions and trends. Moreover, other methodologies such as naive Bayes, maximum entropy classifier, and a linear support vector machine for the sentiment classification of tweets to positive and negative classes have been followed [16]. The above studies exhibited comparable accuracy on their test datasets, but they have certain differences concerning the features they used.

The evolving nature of social networking data has also been tackled in methodologies which demonstrate the temporal dynamics of sentiments in reaction to emerging resources such as a live debate video [4]. Such methodologies offer representation of tweets collections and provide an overall sentiment (positive and negative) micro bloggers understanding. Moreover, using psychology driven methodologies has been followed in efforts like [2] where an extended version of a well established psychometric instrument, the Profile of Mood States (POMS) is used. This instrument measures six individual dimensions of mood (tension, depression, anger, vigour, fatigue, and confusion) to employ a tweets sentiment analysis via a structure of a six-dimensional vector which aggregates mood components on a daily scale comparing their results to the timeline of cultural, social, economic, and political events which took place in that defined period of time.

Dealing with a spectrum of emotional scale is also raised in [21] where eight primary emotion dimensions (acceptance, fear, anger, joy, anticipation, sadness, disgust and surprise) and their synonymous adjectives, are the ones to determine tweets sentiments via a similarity clustering methodology. Such methodologies can be applied by additionally using dictionaries which support semantics and emotional similarities.

4. Frameworks and Applications

Identifying the events that affect social interactions over time is important since many applications can flourish from such activities. For example, by tracking posts in a micro-blogging website someone can identify intensity in emergency situations (floods, fires, riots, etc) or in other events which are of interest to the general public.

Highlighting trends in users' opinions, and revealing their preferences, is of major interest to stakeholders who would like to can track customers' opinions and complaints in a timely fashion such that strategic decisions can be facilitated. Moreover, tracking the evolution of groups (communities) of users or resources, and finding changes and correlations in time can lead to better personalized recommender systems which will improve user experience. Therefore, social data mining and affective analysis results can be useful for popular applications such as recommender systems, automatic event detectors, etc.

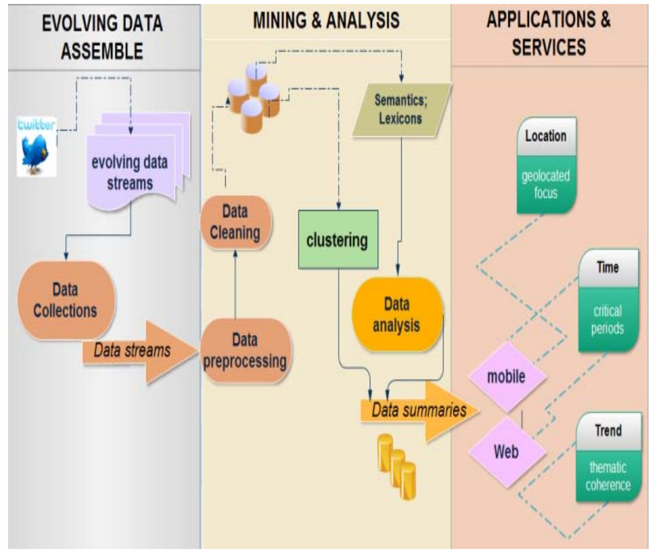

Figure 3. A mining and analysis framework for evolving social data

Figure 3 summarizes a flexible framework which can drive social evolving data mining and analysis applications development. This generic framework separates the following tiers :

- Data assemble : social data streams are gathered typically via the provided social media APIs (twitter is a popular social media as displayed in Figure 3);
- Data mining and analysis : social data streams are refined typically via pre-processing and cleaning techniques in order then to proceed to a mining and or lexicon-based emotional analysis processes and end up with data summaries;
- Applications and services: this tier exploits the produced data summaries in order to develop applications of use in multiple cases. Emphasis should be placed in criteria such as time, location and trend detection since these are the most relevant to the social media audiences.

Next, we focus on two indicative applications developed based on the above framework with a focus on a trend detection and a geo-located emotional crowd pulse capturing.

4.1 Social media trend detection via cloud infrastructure

Cloud4Trends is a microblogging and blogging localized content collection and analysis framework for detecting currently popular topics of users' interest. This framework focuses on collecting and analyzing user generated content through microblogging and blogging applications, both separately and jointly, focused on certain geographical areas, towards the identification of the most significant topics using trend analysis techniques [22]. Here, an application is presented based on the above framework with a focus on evolving social media trend detection on certain geo-located areas. Based on the above structures and mining sections, this application was built by the use of similarity key-value structures which characterize tweets relevance and via a Map-Reduce like clustering methodology.

The application was build on a cloud computing infrastructure to offer a significant benefit in dealing with massive evolving social data sizes produced daily. A detailed system architecture model was also proposed, largely based on a set of service modules developed within the VENUS-C *(Virtual Multidisciplinary EnviroNments USing Cloud Infrastructures, http://www.venus-c.eu/)* research project to facilitate the deployment of research applications on Cloud infrastructures.

This application exploits clustering methodologies and the proposed framework in earlier sections on the basis of the following principles:

- Massive content sizes and unpredictable content generation rates in evolving social networking activities require scalable analysis;
- Trending topics in evolving social data streams should be discovered when they are up to date and an on-line analysis approach is demanded;
- Trends in evolving social media should be meaningful in a particular contextual setting;
- Evolving social content is dispersed in multiple sources and trend detection needs a combined approach

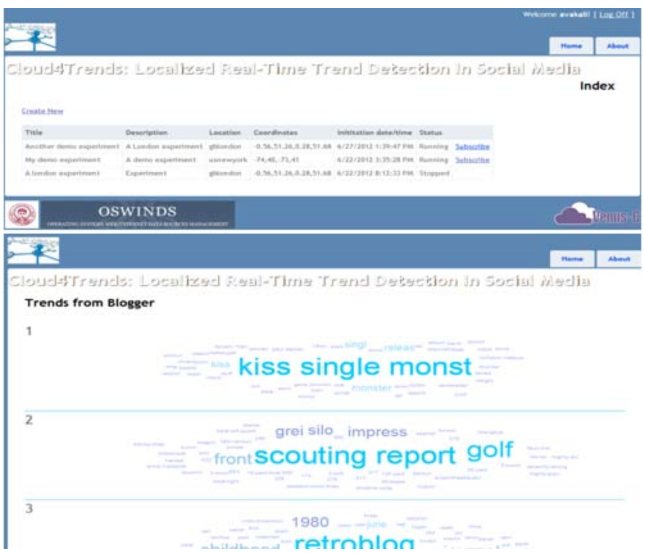

Figure 4. Cloud4Trends application

http://cloud4trendsdemo.cloudapp.net

Figure 4 depicts such an application which sets the experiments by leveraging the cloud infrastructure (top part of Figure 4) and then localized real-time trend detection in social media is discovered and displayed in tag cloud fashion (bottom part of Figure 4). This application exposed the flexibility of the proposed framework and as it was discovered :

- Social media reflect societal concerns exhibiting 'bursts' of content generation on the occurrence of events;
- popular topics / interests fluctuate with time;
- Challenging issues arise for both computer scientists and application developers to reach unbiased, meaningful conclusions about trending users' opinion and interests.

The Cloud deployment of the Cloud4Trends scenario with use of the appropriate clustering methodologies verified that Cloud-based architectures are a viable solution for online social evolving data mining applications that are beneficial for both researchers and entrepreneurs.

4.2 Capturing Emotional spectrum over microblogs

CapturEmos is an application for capturing emotional patterns in micro-blogging data streams via a fine-grained analysis of an emotional spectrum set along with emotion intensity scaling for a particular topic/product, at a particular area and at a particular time period. Based on the above structures and mining sections, this application was built by the use of similarity matrices which characterize tweets sentiment relevance (via a semantic and a sentiment lexicon)[1] and via a scaling mining methodology.

Figure 5. CapturEmos application

http://oswinds2.csd.auth.gr/EmoGlobe/

This application was built on the basis of the framework given in Figure 3 and on the basis of the following principles :

- capturing and understanding crowd's emotions for a particular topic or product in an implicit manner via computational methods;
- sentiment analysis & microblogging (statistical) processing places emphasis on affective and opinion mining, via lexicon-based processing, knowledge extraction techniques;
- the development of relevant Web applications should be enhanced with of crowds emotions visualization capabilities

The innovation principle of this application lies in the fact that focus in evolving social media should be placed on the affect issue

[1] *Wordnet http://wordnet.princeton.edu/ and Sentiwordnet http://sentiwordnet.isti.cnr*

which is distinguished from discrete emotions since these concern affective reactions in relation to one's goals, whereas affect refers to an overarching positive or negative valence of one's feelings.

Figure 5 depicts such an application (called CapturEmos) which works on particular datasets relevant to evolving social media on a particular area and with respect to a particular emotion (set in the interface of the Figure 5 top part) and then localized emotional reaction in a social media microblogging setting (Twitter was used in the example) is discovered and displayed in a map coloring fashion (bottom part of Figure 5).

This application also exposed the flexibility of the proposed framework (Figure 3) and as it was discovered:

- human reactions are very important in evolving social media circulation and emotional spectrum reveals crowd pulse;
- affect analysis captures people's opinions fluctuating on a particular location and in response to a particular thematic context;
- Challenging issues arise for both computer scientists and psychologists to reach meaningful conclusions about users' opinions and behavioral reactions.

5. Conclusions and Future Challenges

This paper unified approaches in mining and affective analysis in order to propose a framework suitable for applications design and development in the evolving social data reality. The appropriate data structures and their relevant methodologies were highlighted and then the proposed design framework encapsulated all tiers needed to support social evolving data cleaning, summarizing and analysis. Two indicative applications were presented as relevant scenarios with a focus on trend detection and emotional crowd capturing. Such applications are useful for both scientists and stakeholders since they place research challenges and they have a large impact in policy and decision making.

Future work should place emphasis on emerging, and unpredicted bursts detections in evolving social media as well as on user multi-profiles patterns since these criteria are crucial in trend and emotion detection. Challenges also originate from the need to support applications with multi-lingual support, and with privacy and anonymity preservation. Finally the real-time emerging nature of the evolving social data flows demands careful design and management in terms of time and space complexities and development of intelligent and collective information retrieval techniques are required and well expected. The proposed framework and applications address wide stakeholders and markets audiences since certain tasks (such as capturing branding success and diffusion in the market, as expressed by the crowds emotions) are important in policy and decision making.

6. ACKNOWLEDGMENTS

The author thanks Dr. Vassiliki Koutsonikola, as well as Maria Giatsoglou and Despoina Chatzakou phd students at the Department of Informatics of the Aristotle University of Thessaloniki, for their support in the presented work and their contribution in the applications development.

7. REFERENCES

[1] Au Yeung, C.M., Gibbins, N., and Shadbolt, N. 2009. Contextualising Tags in Collaborative Tagging Systems. In Proceedings of 20th ACM Conference on Hypertext and Hypermedia, pp. 251-260.

[2] Bollen, J., Pepe, A., Mao, H.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. International Conference on WWW, 2010.

[3] Chi, Y., Zhu, S., Song, X., Tatemura, J., and Tseng, B. L. 2007. Structural and temporal analysis of the blogosphere through community factorization. In Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining. KDD '07. ACM, New York, NY, 163-172.

[4] Diakopoulos, N., Shamma, D.: Characterizing Debate Performance via Aggregated Twitter Sentiment. Proc. of ACM Conf. on Human Factors in Computing Systems (CHI), Atlanta Georgia (2010) 1195--1198

[5] Duan, D., Li, Y., Jin, Y., and Lu, Z. 2009. Community mining on weighted directed graphs. In Proceeding of the 1st ACM international Workshop on Complex Networks Meet information & Knowledge Management . CNIKM '09. ACM, New York, NY, 11-18.

[6] Fortunato, S. & Castellano, C. 2007. Community structure in graphs. arXiv:0712.2716v1.

[7] M. Giatsoglou, A. Vakali, "Capturing Social Data Evolution via Graph Clustering," IEEE Internet Computing, IEEE Computer Society Digital Library, 2012.

[8] Girvan, M. & Newman, M. E. J. 2002. Community structure in social and biological networks. In Proceedings of the National Academy of Sciences of the United States of America, 99(12):7821-7826.

[9] Hotho, A., Robert, J., Christoph, S., and Gerd, S. 2006. Emergent Semantics in BibSonomy. GI Jahrestagung Vol. P-94, 305–312. Gesellschaft fr Informatik.

[10] Koutsonikola, V., Vakali, A., Giannakidou, E., and Kompatsiaris, I. 2009. Clustering of Social Tagging System Users: A Topic and Time Based Approach. In Proceedings of the 10th international Conference on Web information Systems Engineering. G. Vossen, D. D. Long, and J. X. Yu, Eds. Lecture Notes In Computer Science, vol. 5802. Springer-Verlag, Berlin, Heidelberg, 75-86.

[11] Lin, Y., Sundaram, H., Chi, Y., Tatemura, J., and Tseng, B. L. 2007. Blog Community Discovery and Evolution Based on Mutual Awareness Expansion. In Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence. Web Intelligence. IEEE Computer Society, Washington, DC, 48-56.

[12] Liu, B.: Handbook of Natural Language Processing, Second Edition. Nitin Indurkhya, Fred J. Damerau, Goshen, Connecticut, USA (2010)

[13] Mika, P. 2005. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Proceedings of the 4th international SemanticWeb Conference. ISWC'05. Springer Berlin /Heidelberg, pp. 522-536.

[14] H. Ning et al., 2010. Incremental spectral clustering by efficiently updating the eigen-system. Pattern Recogn. 43, 1, 113-127.

[15] O' Connor B., Balasubramanyan R., Routledge B. R. and Smith N. A., "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", Proc. of the

International AAAI conference on Weblogs and Social Media, Washington DC, 2010.

[16] Pak A., Paroubek P., “Twitter as a corpus for Sentiment Analysis and Opinion Mining”, Proc. of the 7th conference on Int Language Resources and Evaluation, 1320-1326, 2010.

[17] Palla, G., Barabási, A.-L., and Vicsek, T. 2007. Quantifying social group evolution. Nature 446, 664-667.

[18] Sun, J., Tao, D., and Faloutsos, C. 2006. Beyond streams and graphs: dynamic tensor analysis. In Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining . KDD '06. ACM, New York, NY, 374-383.

[19] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Association for Computational Linguistics (2011)

[20] H. Tong et al., 2008. Proximity tracking on time-evolving bipartite graphs. SDM‘08. 704–715

[21] Tsagkalidou K., Koutsonikola V., Vakali A., and Kafetsios K., “Emotional aware clustering on micro blogging sources”, In Proceedings of J. Spatial Information Science, pp.387-396, 2011.

[22] A. Vakali, M. Giatsoglou, S. Antaris : “Social Networking Trends and Dynamics Detection via a Cloud-Based Framework Design”, International Conference WWW 2012 , WWW Companion Proceedings, ACM Digital Library, April 2012.

[23] Yang, S., Wu B., Wang, B. 2009. Tracking the Evolution in Social Network: Methods and Results. Complex (1): 693-706.

[24] Zhao, Q., Mitra, P., and Chen, B. 2007. Temporal and information flow based event detection from social text streams. In Proc. of the 22nd National Conference on Artificial intelligence - Volume 2. Aaai Conference On Artificial Intelligence. AAAI Press, 1501-1506.