ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

# Higher-Order Singular Value Decomposition Real - Time Bursty Topic Detection From Twitter

[1]Mrs.C.Senbagavalli, [2] R.Kiruthika
[1]HoD,Department of Information Technology, [2]Research Scholar
[1&2] Kovai Kalaimagal college of Arts and Science,
Coimbatore-641109.

## ABSTRACT

Twitter becomes one of the largest micro blogging platforms for users around the world. These studies have aimed a extracting the period and the location in which a specified topic frequently occurs. Twitter is the most important and timely source from which people find out and track the breaking news before any mainstream media picks up on them and rebroadcast the footage. A bursty topic in Twitter is one that triggers a surge of relevant tweets within a short period of time, which often reflects important events of mass interest. How to leverage Twitter for early detection of bursty topics has therefore become an important research problem with immense practical value. Despite the wealth of research work on topic modelling and analysis in Twitter, it remains a challenge to detect bursty topics in real-time. In this paper, we use framework higher-order singular value decomposition (HOSVD) we focus on a system that detects hot topic in a local area and during a particular period. There can be a variation in the words used even though the posts are essentially about the same hot topic. Topic modelling and analysis in Twitter, it remains a challenge to detect bursty topics in real-time. Our experiments on a large Twitter dataset and synthetic datasets show that the proposed models can effectively mine the topic-specific behavioral factors of users and tweet topics. We further demonstrate that the proposed model consistently outperforms the other state-of-the-art content based models in retweet prediction over time.

Keywords: [Complexity, Rearness, Granularity, Noise, HOSVD]

## 1. INTRODUCTION

The HOSVD is a generalization of the matrix SVD to higher order matrices. Some pioneering and successful applications of the HOSVD in computer vision have been proposed in. In this paper, we demonstrate the aptness of the HOSVD as a transform basis for efficient and effective patch-based denoising. The main thing we wish to emphasize is how such a simple approach yields a performance comparable to techniques that are far more complex conceptually as well from the point of view of implementation. Note that our approach should not be confused with HOSVD-based denoising approaches such as, which are solely designed for hyper spectral images, and which treat the entire image as a single tensor thereby ignoring non-local patch similarity.Content propagates among users through their follow links, from followees to followers. The former are the senders, and the latter are known as the receivers. A receiver may adopt the content exposed to her based on a number of factors, namely the: (a) virality of the sender, (b) susceptibility of the receiver, (c) virality of the content topic, and (d) strength of relationships between sender and receiver. User virality refers to the ability of a user in getting others to propagate her content, while user susceptibility refers to the tendency of a user to adopt her followees' content. Topic virality refers to the tendency of a topic in getting propagated. Since microblogging has been shown rather an information source than a social networking service, we assume in this paper that most relationships among users in a microblogging site are casual and identical in strength. We therefore focus on modeling the user and content factors that drive content propagation without considering the pairwise relationships among users. The modeling of the virality and susceptibility factors has many important applications. In advertisement and marketing, companies may hire viral users to propagate positive content about their products, or to the advertisement with viral content so as to maximize their reach. Similarly, politicians may leverage on viral users to disseminate their messages widely or to conduct campaigning. Also, one may detect events by tracking those mentioned by non susceptible users, and detect rumors based on susceptible users' interactions with the content.

**Complexity:** The computational complexity of finding frequent sequential patterns is huge for large symbol sets. Many existing algorithms have a time complexity that grows exponentially with decreasing pattern supports. **Rareness:** In general, the support of a specific sequential pattern decreases significantly with the growing cardinality. To see this, let us consider k symbols that appear with uniform probability in a sequence. The possibility of locating a particular pattern of length ` is ` −k . In other words, the higher the cardinality, the rarer the patterns are. Since the number of unique subsequences grows with the cardinality, the number of sequences required to identify significant patterns also tends to grow drastically. **Granularity:** A large number of symbols in a sequence can "dilute" useful patterns which themselves exist at a different level of granularity. As we will discuss in more detail later, semantically meaningful patterns can exist at a higher granularity level, therefore pattern mining on the original, huge set of symbols may provide few clues on interesting temporal structures. **Noise:** Due to the stochastic nature of many practical sequential events, or the multi-modality of events, useful patterns do not always repeat exactly but instead can happen in many permutations. For example, the customers may accidentally download some trial products by mistake when they are looking for the desired information. Without dealing with such irregular perturbations, we may fail to discover some meaningful patterns.

## 2. LITERATURE REVIEW

**Amr Ahmed** et al show how this model can be applied to data from a major Internet News portal. From the view of statistics, topic models, such as Latent Dirichlet Allocation (LDA), and clustering serve two rather incomparable goals, both of which are suitable to address the above problems partially. **Blei et al.** provide insight into the content of documents by exploiting exchangeability rather than independence when modeling documents. **Li and McCallum &** Teh et al. related to Pachinko Allocation and the Hierarchical Dirichlet Process. **Griffiths and Steyvers** has been previous work on scalable inference, starting with the collapsed sampler representation for LDA and efficient sampling algorithms that exploit sparsity. **Canini et al** address this by designing an SMC sampler which is executed in parallel by allocating particles to cores. The data structure is a variant of the tree described. **Pearson,** has proved invaluable in a number of application domains. The basic paradigm is simple and intuitive: (i) compute certain statistics of the data—often empirical moments such as means and correlations—and (ii) find model parameters that give rise to (nearly) the same corresponding population quantities. **Le Cam** viewed as complementary to the maximum likelihood approach; simply taking a single step of Newton-Raphson on the likelihood function starting from the moment based estimator. **Roch, Hsu et al, Anandkumar et al, Hsu and Kakade** make the decomposition explicit under a unified framework. Specifically, we express the observable moments as sums of rank-one terms, and reduce the parameter estimation task to the problem of extracting a symmetric orthogonal decomposition of a symmetric tensor derived from these observable moments. **Lathauwer et al.** proposed that convergence analysis of this method for orthogonally decomposable symmetric tensors, as well as a detailed perturbation analysis for a robust (and a computationally tractable) variant. **Baeza-Yates and Ribeiro-Neto,** proposed that find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments. Significant progress has been made on this problem by researchers in the field of information retrieval (IR).

## 3. PROBLEM SPECIFICATION

This real-time task is challenging for existing algorithms because of the high computational complexity inherent in the topic models as well as the ways in which the topics are usually learnt, e.g., Gibbs Sampling or variational inference. The key research challenge is how to solve the following two problems in real-time: (I) How to efficiently maintain proper statistics to trigger detection; and (II) How to model bursty topics without the chance to examine the entire set of relevant tweets as in traditional topic modeling. While some work such as indeed detects events in real-time, it requires pre-defined keywords for the topics.

## 4. OBJECTIVE OF THE RESEARCH

- ➢ Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
- ➢ It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
- ➢ When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

## 5. CONTRIBUTIONS

- Efficiently maintains at a low computational cost the acceleration of two quantities
- These accelerations provide as early as possible the indicators of a potential surge of tweet popularity.
- Update these statistics efficiently and invoke the more computationally expensive topic inference part only
- It possible to achieve real-time detection in a data stream of Twitter scale.
- Our solution can detect bursty topics in real-time, and present them in finer-granularity.

- Variety of difficulties in mining sequential patterns from massive data represented by a huge set of symbolic features.
- Reduces the representation of the sequential data by uncovering significant, hidden temporal structures.

## 6. PROPOSED METHODS

A bursty topic in Twitter is one that triggers a surge of relevant tweets within a short period of time, which often reflects important events of mass interest. How to leverage Twitter for early detection of bursty topics has therefore become an important research problem with immense practical value. Despite the wealth of research work on topic modeling and analysis in Twitter, it remains a challenge to detect bursty topics in real-time. In this paper, we use framework higher-order singular value decomposition (HOSVD) we focus on a system that detects hot topic in a local area and during a particular period. There can be a variation in the words used even though the posts are essentially about the same hot topic. Topic modeling and analysis in Twitter, it remains a challenge to detect bursty topics in real-time. Our experiments on a large Twitter dataset and synthetic datasets show that the proposed models can effectively mine the topic-specific behavioral factors of users and tweet topics. We further demonstrate that the proposed model consistently outperforms the other state-of-the-art content based models in retweet prediction over time.

### 6.1 HIGHER-ORDER SINGULAR VALUE DECOMPOSITION (HOSVD)

Singular value decomposition (SVD) has been studied and used extensively in diverse fields. Its properties, such as uniqueness, best low-rank approximation, etc, lend itself to many applications in scientific computing; e.g. image analysis, clustering, to name a few. Recently an increasing number of problems in different domains involve manipulations of higher dimensional data, and tensors are a natural generalization of matrices into higher dimensions. An order N tensor is simply a N dimensional array in which each element is indexed by a N-tuple (i1, i2, . . . , iN ). Due to the usefulness of matrix SVD, it is natural to look for a similar higher order decomposition with similar properties. In this report, we present a proper generalization of such higher order SVD and discuss some of its properties in relation to the normal matrix SVD. It is worth mentioning that other SVD generalizations exist, and one could choose whichever to use based on which matrix SVD property one wants to carry over to higher dimension. The HOSVD we present here is probably one bearing the most resemblance in decomposition form to matrix SVD, and also possess strikingly analogous properties. The report is organized as follows. Section 3 introduces the necessary definitions of some higher order operations analogous to the operations on matrices, and states the main theorem of HOSVD.

Definition 1. Let A € ₵ $^{I1xI2X\ldots x\ IN}$ be an order N tensor. The n-mode matrix unfolding (also called matricization) $A_{(n)}$ € ₵ $^{In}$ $_{x(I(n+1x\ldots xINxI1x\ldots In-1)}$ has tensor element $a_{i1,i2,\ldots iN}$ at index $(i_n,j)$ where,

$$j = 1 + \sum_{k=1,k\neq n}^{N} (ik - 1)J_k \quad \text{with} \quad J_k = \prod_{m=1,m\neq n}^{N} I_m$$

**Definition 2.** The n-mode product of tensor A € ₵ $^{I1xI2X\ldots x\ IN}$ with matrix U € ₵ $^{In\ x\ J}$ is of size $I_1$ x … x $I_{n-1}$ x J x $I_{n+1}$ x … x $I_{N,}$ and defined element-wise as

$$(A\ x_n\ U)_{i1\ldots\ in-1\ jn+1\ldots iN} = \sum_{in=1}^{IN} ai1i2 \ldots iN\ Ujin$$

A better characterization is that each mode-n column vector (by fixing indices for all other modes) is multiplied by matrix U; in matricized form we have

$y = A \, x_n \, U \iff y_{(n)} = UA_{(n)}$ (1)

$A \, (I_1 \times I_2 \times \ldots \times I_N)$ – tensor A can be written as the product

$A = S \times_1 U^{(1)} \times_2 U^{(2)} \ldots \times_N U^{(N)}$ (2)

(a) All-orthogonality: for all n, substensors $S_{in=α}$ and $S_{in=ß}$ are orthogonal for all α and ß with α≠ß.

(b) Ordering: for all n,

$||S_{in=1}|| \geq ||S_{in=2}|| \geq \ldots \geq ||S_{in=In}|| \geq 0$ (3)

The Frobenius norm $||S_{in=i}||$ , denoted by $σ^{(n)}_{i,}$ are the n-mode singular values of A and the columns of $U^{(n)}_i$ are the n-mode singular vectors.

We derive the algorithm to actually compute HOSVD. It is indeed a straightforward consequence of Theorem 1. By equation (2) and the commutative property of n-mode product, we can easily obtain the following relation

$S = A \, X_1 \, U^{(1)H} \ldots X_N \, U^{(N)H}$ (4)

Repeated application of the relation (1), together with the definitions of Kronecker product leads to the n-mode unfolding of A as

$A_{(n)} = U^{(n)}S_{(n)}(U^{(n+1)}$ ¤ … ¤ $U^{(N)}$ ¤ $U^{(1)}$ ¤ … ¤ $(U^{(n-1)})^H$ (5)

The all - orthogonality and ordering properties of S imply that $S_{(n)}$ has mutually orthogonal rows with Frobenius norms equal to $σ^{(n)}_1, σ^{(n)}_2 ,\ldots, σ^{(n)}_{In.}$

Notice that since each $U^{(n)}$ is orthogonal, by the definition of Kronecker product,

$U^{(n+1)}$ ¤ … ¤ $U^{(N)}$ ¤ $U^{(1)}$ ¤ … ¤ $(U^{(n-1)}$

is also orthogonal. Therefore, we define

$Σ^{(n)} = \mathrm{diag}(σ^{(n)}_1, σ^{(n)}_2 ,\ldots, σ^{(n)}_{In})$ (6)

and columnwise orthonormal V as

$V^{(n)} = (U^{(n+1)}$ ¤ … ¤ $U^{(N)}$ ¤ $U^{(1)}$ ¤ … ¤ $(U^{(n-1)})\hat{S}^H_{(n)}$ (7)

where $\hat{S}_{(n)}$ is normalized version of $S_{(n)}$ with each row scaled to unit norm (i.e. $S_{(n)} = Σ^{(n)} \hat{S}_{(n)}$)

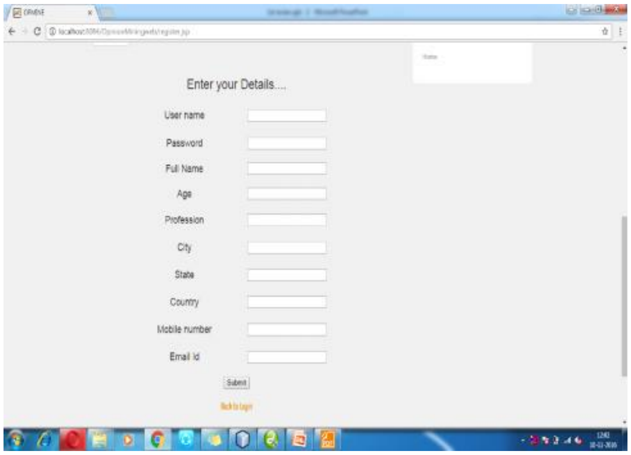
$A_{(n)} = U^{(n)} \, Σ^{(n)} \, V^{(n) \, H}$ (8)

Equations (4) and (8) lead to the following intuitive and straightforward algorithm outlined mentioned below.

**Algorithm 1** Procedure for computing HOSVD

```
1: for n = 1,…,N do
    2: Compute $A_{(n)} = U^{(n)} Σ^{(n)} V^{(n) H}$
3: end for
4: $S = A \times_1 U^{(1)H} \ldots \times_N U^{(N)H}$
5: return $S, U^{(1)}, \ldots, U^{(N)}$
```

## 7. RESULT



Enter your Details....

User name

Password

Full Name

Age

Profession

City

State

Country

Mobile number

Email Id

Submit

Back to login

**Figure 1 :Enter Details**

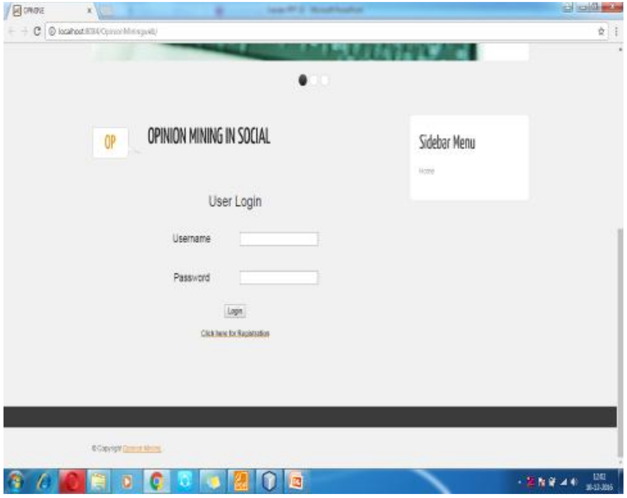

OPINIONE
localhost:8084/OpinionMiningweb/
OP
OPINION MINING IN SOCIAL
Sidebar Menu
Home
User Login
Username
Password
Login
Click here for Registration
©Copyright Opinion Mining

Figure 2:User Login

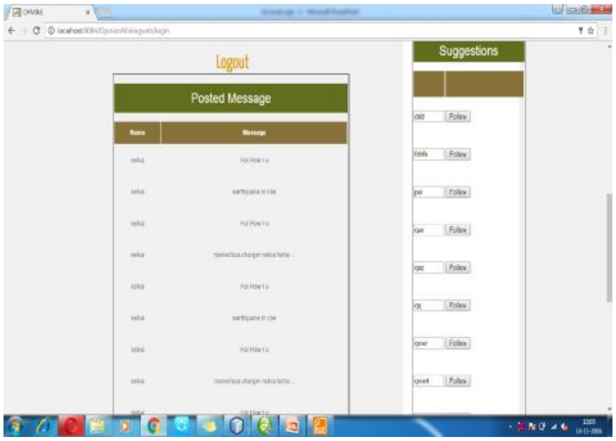


Figure 3: Logout

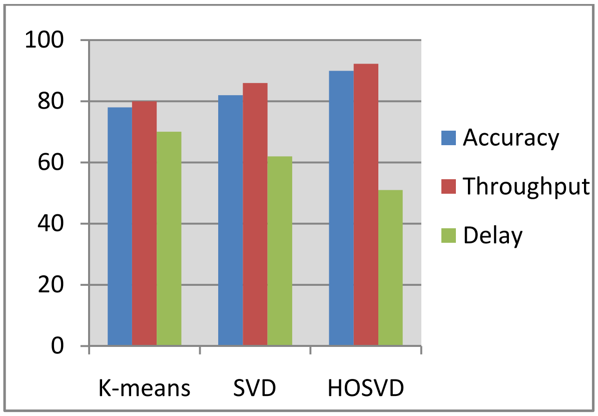


Figure 4 :Performance Graph

This Figure 4 has shown that fairly simple techniques can achieve very high quality results, but that substantial work is needed to reduce the errors to manageable numbers. Fortunately, that the problem focuses on Broadcast News and not on arbitrary forms of information means that there is hope that more carefully crafted approaches can improve the tracking results substantially.

## 8. CONCLUSION AND FUTURE WORK

We study user and content factors underlying use framework higher-order singular value decomposition (HOSVD) we focus on a system that detects hot topic in a local area and during a particular period. There can be a variation in the words used even though the posts are essentially about the same hot topic. Topic modelling and analysis in Twitter, it remains a challenge to detect bursty topics in real-time. Our experiments on a large Twitter dataset and synthetic datasets show that the proposed models can effectively mine the topic-specific behavioral factors of users and tweet topics. We further demonstrate that the proposed model consistently outperforms the other state-of-the-art content based models in retweet prediction over time. We proposed dimension reduction techniques based on hashing to achieve scalability and, at the same time, maintain topic quality with robustness. We also presented case studies on interesting bursty topic examples which illustrate some desirable features of our approach. While in this work, we provide more sophisticated results, which capture not only the information of word pairs, but also the word triples; more effective inference algorithm, i.e. HOSVD, which is an important contribution and more comprehensive evaluations.

## REFERENCES

[1] A. Ahmed, Q. Ho, C. H. Teo, J. Eisenstein, A. J. Smola, and E. P. Xing. Online inference for the infinite topic-cluster model: Storylines from streaming text. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics,

AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011, pages 101–109, 2011.

[2] J. Allan, R. Papka, and V. Lavrenko. Online new event detection and tracking. In SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, pages 37–45, 1998.

[3] F. Alvanaki, S. Michel, K. Ramamritham, and G. Weikum. See what's enblogue: real-time emergent topic identification in social media. In 15th International Conference on Extending

Database Technology, EDBT '12, Berlin, Germany, March 27-30, 2012, Proceedings, pages 336–347, 2012.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. Journal of Machine Learning Research, 15(1):2773–2832, 2014.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. The Journal of machine Learning research, 3:993–1022, 2003.

[6] D. M. Blei and J. D. Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113–120, 2006.

[7] T. Brants and F. Chen. A system for new event detection. In SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada, pages 330–337, 2003.

[8] M. Cataldi, L. D. Caro, and C. Schifanella. Personalized emerging topic detection based on a term aging model. ACM TIST, 5(1):7, 2013.

[9] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In Advances in Neural Information Processing Systems 22: 23rd

Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada., pages 288–296, 2009.

[10] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. Journal of Algorithms, 55(1):58–75, 2005.

[11] Q. Diao, J. Jiang, F. Zhu, and E. Lim. Finding bursty topics from microblogs. In The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers, pages 536–544, 2012.

[12] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In Proceedings of the 21th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pages 219–228, 2015.

[13] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In 31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015, pages 1561–1572, 2015.

[14] T. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1):5228–5235, 2004.

[15] P. Guttorp. An introduction to the theory of point processes (D. j. daley and d. vere-jones). SIAM Review, 32(1):175–176, 1990.

[16] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. Biometrika, 58(1):83–90, 1971.

[17] D. He and D. S. P. Jr. Topic dynamics: an alternative model of bursts in streams of topics. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010, pages 443–452, 2010.

[18] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999.