International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

Data Mining Using Matrix Factorization for Enhancing a Patient's HealthCare

[1] Manigandan. J [2] Dr. Soundararajan. S, M.E., PhD,
[1] M.E. Computer Science and Engineering, [2] Vice Principal
[1][2] Velammal Institute of Technology,
Panchetti, Chennai.
[1] rjsmanigandan23@gmail.com. [2] viceprincipalvit@gmail.com.

Abstract: **Web mining is used to discover as well as extract data from web documents and service. Social networking sites are used to discuss the current topics and reactions to current happening on the internet. The discussion which reflects the opinion of people, thoughts and their innovative ideas. Detection of current topic and tracking valid data from offline articles is quite difficult. Detection of topic from social networking sites will helps to gather and analyses the huge volume of up-to-the minute. Topics are detected based on vigorously and provides path to various treatments to cure the diseases. The techniques are called as Formal Concept Analysis [3] based on Matrix Factorization are intended to pick up the evolution and issues of current topic in unstructured content which are present in a social media. Extraction and analyses of data based on the user-needed data content. Self organizing maps [16] are used to correlate the data based on positive and negative words present in the user's status. Scores of text will give as numerical value of each user forums. The pictorial representation can be viewed based on the scored values and for easy understanding. It helps to determine the better treatments and least cost medicine to cure incurable diseases can be identified and try to cure by early stage as soon as possible.**

Index Terms: Cluster Analsyis, FCA, Grid factorization, Neural Networks, NMF, R Console, SOM, Twitter.

I. INTRODUCTION

The Internet has become the major part in day to day life of Kinsfolk. Nobody will imagine a world without internet. Were each and everyone getting to grasp things that happen around the world through internet? This project is to detect disease and various diseases from the foremost noted social networking web site known as Twitter that exhibits the views of individuals on certain person, event, and organization. Web Usage Mining[11] which helps out to explore interesting web usage facts in order to understand the needs of the user in which tends help or provide the needy to users who use the web applications. Web usage mining is a mechanism of extracting useful facts from server activity reports.

The count of social networking sites users will reach 2.55 billion by the end of the year 2017 this proves that how people show their interest in sharing their views and opinions in social networking sites such as Orkut, Twitter, Face book, and Google plus. This project aims out to detect various treatments for incurable disease like lung cancer, blood caner, etc., from Twitter. Multiple treatments for diseases detection is a seminal event or activity which is directly related to disease. Multiple treatments for diseases detection helps out in knowing the most talked about and important happenings that has been spoken all over the social networking sites. All these social contents are analyzed to detect the multiple treatments for diseases from the posts and links such as link mining [15], classification through links [15], predictions based on objects as well as links [8], existence [10], estimation [11], object [7], group and subgroup detection [12], and mining the data [4], [8].
Link prediction, viral marketing [13], and online discussion [4][14] groups (and rankings) allow for the development of solutions based on user feedback. When the social contents are analyzed with appropriate statistical and computational tools, social media contents can be turned into invaluable future Insights.

Detection of incurable disease for multiple treatments which helps out to know how public reacted to a particular cured from various disease and recover their lives. And also it helps us to know how new medicine and current booming medicine get evolved and who has been influenced because of the news can also be retrieved. Healthcare providers could use patient opinion to improve their services. Physicians could collect feedback from other doctors and patients to improve their treatment recommendations and results.

II. PROBLEM DEFINITION

A. Existing System

In the web mining scenario, the records to match are highly query-dependent, since they can only be obtained through online forums. Moreover, the dictionary has capability to store only 110 words i.e., 55 positive and 55 negative words generally. And also, the dictionary is not efficient to analyses the medical terms and their semantic definitions.

B. Problem Statement

The existing research does not sufficiently assess whether sentiments can be analyzed using sentiment function. And by scoring the tweets using positive and negative scores within the limited number of 110 words are stored in dictionary. The dictionary is not efficient to analyses a medicinal terms and semantically words of medicines.

C. Objectives of the Proposed System

The aim of the work is more clearly to retrieve and analyze the sentiments of a cancer and incurable disease from twitter posts. In order, to add more medical termed words to dictionary. Efficient to retrieve the medical terms and dictionary words. Determine the least cost treatments to cure the cancer diseases.

III. EXTRACTION OF TWEETS

A. Create an Application on Twitter

Twitter helps out the students and research scholars to do something useful with tweets. There are certain steps and procedures to fetch tweets from the twitter. The first step is to create an application [4] in twitter which asks for username and password of the creators twitter account. Once the application is created successfully twitter generates a consumer key and a consumer secret key which is like a basic gateway in fetching out tweets from twitter.

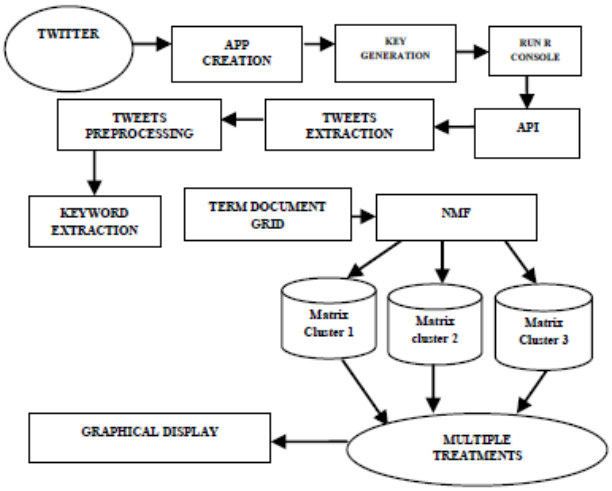

Fig.1. Functional Architecture for detection of diseases And their multiple treatments from Twitter.

B. Execution of R console

The main purpose of using R console is R is an open source environment for statistical computing and graphics. Another major reason of using R is it provides a wide variety of statistical and graphical techniques. R can retrieve tweets from twitter by running its chunk of code one and only if twitter supporting packages are installed such as twitter, ROAuth, plyr, stringr, ggplot2. Once these packages are installed the consumer key and the consumer secret is entered in the code which was given by the twitter when the new application was created. After the execution of these lines of code a API link will be generated which was shown in Fig.1. After the link is copied and pasted in twitter a 7 – digit pin is generated which will be source to retrieve tweets. Pin generation and execution of that pin will create handshake based authentication which acts as a pathway between R and Twitter. In twitter, the nodes are interconnected and they form a clustering of nodes.

Similar nodes are interconnected with each other to retrieve a user needs data. Self-Organizing Map (SOM) [16] will helps to organize the data based on the links of each other from twitter. Using Apirori algorithm we can capable to retrieve the data and subsets of data from social network.

Apirori Algorithm:

C_k: Candidate itemsets of size k
L_k : frequent itemsets of size k
L_1 = {frequent 1-itemsets};
for (k = 2; L_k !=Æ; k++)
C_{k+1} = GenerateCandidates(L_k)
For each transaction **t** in database do increment count of candidates in
C_{k+1} that are contained in **t**
endfor
L_{k+1} = candidates in C_{k+1} with support
≥**min_sup**
endfor
return È$_k$ L_k**;**

C. Hash Based Tweets

Hash based Tweets is nothing but the type or it can called as calling tweets like for example if tweets related to cancer is to be downloaded from twitter it should be called as #blood cancer were a hash(#) symbol must be priory added before the word were the tweets about is to be retrieved from twitter.

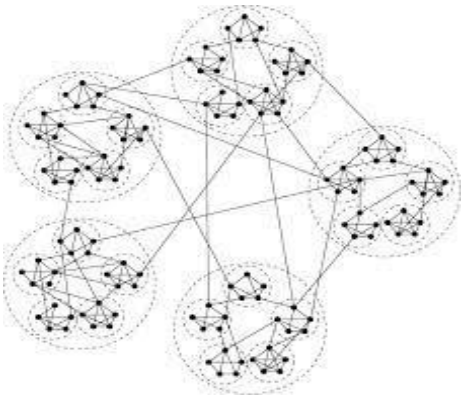

Fig. 2. Cluster based web mining of data from Twitter

IV. PREPROCESSING OF TWEETS

As explained in the Fig.1. the functional architecture of this paper goes like this were after the API link is generated in R console by running its R chunks it generates a 7 digit pin by entering that pin and by executing the code the tweets can be downloaded based their hash tag(#). Once the tweets are downloaded and saved as CSV files and immediately if the tweets are analyzed several noise data and unwanted symbols appear that are not necessary. Once the tweets are downloaded the symbols are first removed from the twitter using java code and the process goes like this by removing special symbols and filling space in the places were symbols previously existed.

Once the symbols are removed from the tweets and filled with the space they are ready for further smoothing process.

We then assigned weights for each of the words found in the user posts using with the following formula:

$$\text{Weight}_{i,d} \begin{cases} \text{Log}\,(tf_{i,d} + 1)\log n\, x/x_i & \text{if } tf_{i,d} > 1 \\ 0 & \text{otherwise} \end{cases}$$

In which tf_i, d represents the word frequency (t) in the document (d), n represents the number of documents within the entire collection, and xt represents the number of documents where t occurs. Once the symbols are removed from the tweets and filled Times New Roman with the space they are ready for further smoothing process.

A. Text Transformation

Once the symbols are removed the tweets are first transformed to a data frame which is used for storing data tables. They are nothing but list of common size vectors.

```
requestURL = "https://api.twitter.com/oauth/request_token"
accessURL = "https://api.twitter.com/oauth/access_token"
authURL = "https://api.twitter.com/oauth/authorize"
consumerKey = "xxxxxxxxxxxxxxxxxxxxx"
consumerSecret = "xxxxxxxxxxxxxxxxxxxxxxxxxx"
Cred <- OAuthFactory$new(consumerKey=consumerKey,
consumerSecret=consumerSecret,requestURL=requestURL,
accessURL=accessURL, authURL=authURL)
Cred$handshake(cainfo = system.file("CurlSSL", "cacert.pem",
package = "RCurl") )
XXXXXX
save(Cred, file="twitter authentication.Rdata")
registerTwitterOAuth(Cred)
Hashtag.list <- searchTwitter('#Hashtag', n=1000,
cainfo="cacert.pem")
Hashtag.df = twListToDF(Hashtag.list)
write.csv(Hashtag.df, file='C:/temp/HashtagTweets.csv',
row.names=F)
```

Fig. 3. R Chunk to Download Tweets from Twitter

The built in data frames are used to create or convert data frames from tweets. The data frames are like primary data structure in R which is used to exhibit data. A data frame is a table which can be also called as two-dimensional array-like structure, were each column contains measurements on one variable, and each row contains a case.

There are no constraints in R that all columns in a table should have same features one column may be of numerical other may be of character. After converting the tweets into data frames then they are converted to data corpus mainly for the purpose of specifying the source to be of character vectors. Once the corpus is created there are certain transformations need to be made they are changing upper case letters into lower case letters and removing numbers, stop words, and hyperlinks are also will be removed.

B. Stemming of Words and text

To perform the above mentioned operations such as transforming text and to stem words from the downloaded tweets tm package from library must be installed without that no operations can be performed. In most of the applications the words needed to be stemmed to retrieve their radicals, so that several forms can be retrieved from a stem that can be taken as same word while determining their frequency in occurrence.

Hence the tweets are preprocessed finally by the completion of stop words removal and stemming of words.

A term document grid [9] represents the relationship between terms and documents where each row stands for terms and each column stands for documents.

Positive	Negative	Positive	Negative
Appreciate	Cannot	Agree	Bad
Beneficial	Concern	Benefit	Loss
Importance	Damage	Ease	Don't
Effective	Didn't	Easier	Died
Greatly	Difficult	Good	Isn' t
Favorably	Depression	feasible	Doubt
Favorable	Error	Great	Failure
Grateful	Impossible	Help	Hasn't
Greatest	Hasn't	Enjoy	Hate
Greater	Hurt	Helped	Fear
Helpful	Discomfort	Hoped	Lack
Hopeful	Limited	Honest	Lose
Honestly	Concerns	Hope	Lost
Helping	Miss	Helps	Nasty
Hopefully	Nausea	Hopes	Negative
Hoping	No	Comfort	No
Importantly	Painful	Like	Poor
Improve	Problem	Improved	Sad
Improvement	problems	Improves	scared
Improving	Scary	Inspiration	severe
Impresses	Sorry	Love	hate

Table I. General stored dictionary words.

V. BUILDING A TERM DOCUMENT GRID

The term document grid[9] is formed from the above processed data corpus. Term document grid can be build in R using Term Document Grid[9] function. Since the multiple treatments for diseases is being detected from the tweets and the csv file consists of about 1000 tweets.

Tweets	Tweet 1	Tweet 2	Tweet 3	Tweet 4	Tweet 5
Term 1	1	3	7	4	2
Term 2	4	8	5	9	6
Term 3	3	5	2	1	4
Term 4	1	3	6	2	1
Term 5	2	6	7	3	4

Table II: Term Document Grid

So each and every single tweet is to be considered as a single document and the term document grid will be formed in such a manner that all keywords found in all 1000 tweets will be mentioned in the rows respectively as shown in Table I, the tweets are the documents and terms are the rows. Table I. illustrates that term 1 has been appeared 1 time in document 1 i.e. in tweet 1 and similarly term 2 has been appeared 8 times in tweet 2 respectively. All the numerical value signifies that the frequency of all terms in each and every document i.e. each and every tweet. And hence finally a perfect term document grid[9] is build successfully.

A. Mining of medical termed text and Pre-processing

A Rapid miner (www.rapidminer.com) data collection [8] and processing tree was developed to look for the most common positive and negative words and their term-frequency-inverse document frequency (TFIDF) [13] [1]scores within each post.

Acne	Headache	Itchin	Cachexia	Pneumonia
Throat cancer	Stomach cancer	Zika virus	Viral latency	Chickenpox
vomiting	Genetic disease	Blood cancer	Lung cancer	weakness
Diabetes	bacteria	Rash	Heart disease	Viral disease

Table III. Dictionary stored medical termed words

To this goal, we used the National Library of Medicine's Medical Sub-ject Heading (MeSH), which is controlled vocabulary. (http://www.nlm.nih.gov/mesh/) that consists of a hierarchy of descriptors and qualifiers that are used to annotate medical terms.

A custom designed program was used to map words in the forum to the MeSH database. A list of words present in forum posts that were associated to treatment side effects was thus compiled. This was done by selecting the words simultaneously annotated with a specific list of qualifiers in MeSH (CI– chemically induced; CO – complications; DI – diagnosis; PA– pathology, and PP – physiopathology)

B. Data score function

The score function of tweet based on the medical terms and semantics words of dictionary stored in it. TFIDF [1], [13] are used to correlation of user posts and forums found in each modules. They help us to analysis the positive and negative terms of words are known as Module Average Opinion (MAO)[1] by equation written as

$$MAO = \frac{Sum_{+} - Sum_{-}}{Sum_{all}}$$

$Sum_{+} = x_{ij}$ is the total sum of the TF-IDF [1]
, [13] scores will compare the positive words in the Wordlist vectors within the module containing medical Termed stored words in dictionary. The units i represent post index. The unit j represents the wordlist index (matching the positive words in the module list). $Sum_{-} = x_{ij}$ is the total sum of the TF-IDF[13],[1] scores will compare the negative words in the wordlist vectors within the module containing medical termed stored words in dictionary. The units i represent post index. The unit j represents the wordlist index (matching the negative words in the module list).

$$\text{Sum}_{all} = \sum_{i=1}^{N} \sum_{k=1}^{M} x_{ik}$$

Is the sum of the total words in the wordlist vectors within the module containing medical termed stored words in the dictionary?

VI. GRID FACTORIZATION

In linear algebra, a grid factorization [6] is deriving products of matrices from single grid. To reduce the dimensionality of a larger complexity grid into a lesser complexity grid and reduction of dimensionality is the main aim of grid factorization [9].

A. Formal Concept Analysis

Formal Concept Analysis (FCA) [10] is the most effective data analysis methodology which was based on ordered lattice theory. It is also called as concept hierarchical structure representing relationships and attributes in to a particular domain. The main aim of FCA [3] is to define the unit of two parts extension and intension. FCA methodology [7] is used to mine association rules from web usage lattice that has been constructed from web lattice and it has been discovered from those web lattices is used to detect multiple treatments for diseases using FCA[10] techniques called Singular Value Decomposition (SVD) and Nonnegative Grid Factorization (NMF)[9].

Terms	NMF VALUE
Term 1	8.2574
Term 2	14.3030
Term 3	6.4969
Term 4	6.4236
Term 5	10.4298

Table IV. Illustrates the NMF values for the Term Document Grid [9] Table I.

Clearly signifies that the Term 2 has the highest value 14.3030 since the terms occurrence or frequency is comparatively higher than the other terms this may look easy for a 5*5 grid and guess which term has the highest value but if the size of the term document grid is
500*1000 then it cannot be guessed just by seeing here comes the job of NMF [9] to find the multiple treatments for diseases. Hence the multiple treatments for diseases are concept related to Term 2 is the multiple treatments for diseases.

VII.CLUSTERING THE MATRIX

Matrix is being clustered using a formal concept analysis approach called NMF [9] which is the most efficient matrix factorization algorithm. What it does is reduce the dimensionality of the matrix and further produce several matrixes from the term document matrix that has been generated in the earlier phase. There are various types of matrix factorization techniques for clustering they are,

A. NMF

When the input data is non negative, and it restricts F and G to be nonnegative. The standard NMF [6] can be written as, NMF: X+ ≈ F+G+ Using an intuitive notation for X, F, G ≥ 0

B. SEMI-NMF

When the input data has mixed signs, it can restrict G to be nonnegative while placing no restriction on the signs of F. This is called SEMI- NMF. SEMI-NMF: X± ≈ F±G+ Semi-NMF [9] can be motivated by K-means clustering. Let F = (**f**1, …… ,**f**k) be the cluster centroids obtained via K-means clustering. Let G be the cluster indicators: i.e., g k_i =1 if $\mathbf{x}_i$ belongs to cluster ck; g k_i = 0 otherwise.

C. CONVEX-NMF

In general, the basis vectors F= (**f**1, ⋯,**f**k) can be anything in a large space, in particular, a space that contains the space spanned by the columns of X= (**x**1, · · **x**n). Inorder for the vectors F to capture the notion of cluster centroids. It restricts them to lie within the space spanned by the columns of X, i.e., **F**l= w1l**x**1+· · ·+wnl**x**n = X**w**l, or F= XW. Furthermore, **f**l is restricted as a convex combination of the data points. Hence they are called as restricted form of factorization as Convex-NMF [6] and it applies to both nonnegative and mixed-sign input data.

D. TRI-FACTORIZATION

This technique is used to simultaneously cluster the rows and the columns of the input data matrix X, They are considered as the following nonnegative 3 factor decomposition, X ≈ FSGT Note that provides additional degrees of freedom such that the low-rank matrix representation remains accurate while row clusters and *G* gives column clusters. An important special case is that the input X contains a matrix of pair wise similarities: X = XT =W. In this case, F = G= H. The optimization of the symmetric NMF is done.

E. KERNEL NMF

Consider a mapping **x**i →f(**x**i), or X →f(X) = (f(**x**1), · · · ,f(**x**n)).A standard NMF or Semi-NMF like f(X) ≈ FGT would be difficult since F,G will depends explicitly on the mapping function f(·). However, Convex-NMF provides a nice possibility: $\varphi(X) \approx \varphi(X)WG$
Depends only on the kernel K= f T(X)f(X). This kernel extension of NMF [9] is similar to kernel-PCA and kernel K-means.

F. Word Cloud and Graphical display

A word cloud depicts the visual representation of keywords based on their importance as well as values in this paper the highest NMF [9] value holding term will acquire the center and larger part of word cloud respectively the lesser terms with lesser NMF value will hold the remaining left parts with respect to their NMF[9] values. Fig.3.Represents an report about Cancer disease and their causes

Fig. 3. Word Cloud of CANCER.

From 2010 to 2025 describes cancer statistics derived from tweets and their respective scores. One of the reasons for cancer diseases are hypertension prevalence is shown as bar chat statics of men and women Fig. 4 respectively. As a consequence, the estimated number of yearly cancer deaths is expected to increase slightly for men and more for women. With respect to prevalence of hypertension, female mortality is expected to be highest for established market economies and former socialist economies. Female is quite larger compared to males as per derived tweets of cancer statistics.

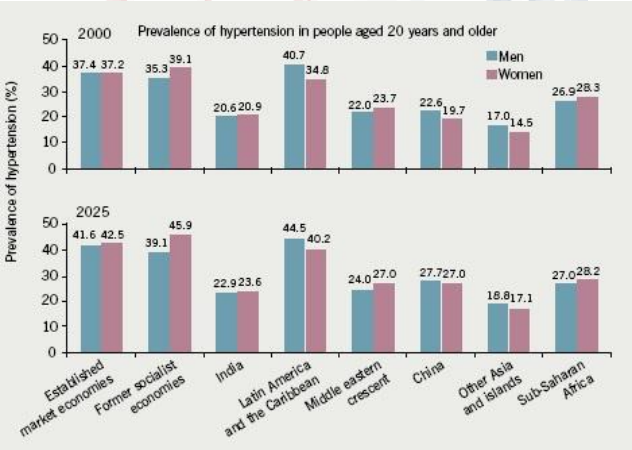

Fig. 4. Bar Chart for CANCER survey from 2000 to 2025.

VIII. CONCLUSION

The main aim of this paper is to detect multiple treatments for diseases from the social networking site twitter using a formal concept analysis based Grid factorization i.e. Non-negative Grid Factorization[9] a best grid factorization technique that can used to determine the current multiple treatments for diseases or new treatments or it can be also called as event detection. Hence this paper proposed a most effective solution and accurate solution in detecting current multiple treatments for diseases. The multiple treatments for diseases are detected on the basis of closest similarity based distance calculation using Nonnegative grid factorization [9] and the multiple treatments for diseases that have been detected will be of perfect accuracy in medical field.

This paper deals with only textual information and now a day's most of the posts in social networking sites are images and videos rather than textual posts. Since the main purpose of inventing NMF [6] is to work with images. So, in future image processing concepts can be built in here to find out multiple treatments for diseases from the images that are being posted and shared in social networking sites [4][12]. Social media can open the door for the health care sector in address cost reduction, product and service optimization, and patient care.

REFERENCES

[1] Altug Akay, Member, IEEE, Andrei Dragomir, Bjorn Erik Erlandsson, Senior Member, *IEEE*, "Network-Based Modeling and Intelligent Dapta Mining of Social Media for Improving Care" 2015.

[2] Ping Li, Jiajun Bu, Yi Yang and RongrongJi, "Discriminative Orthogonal Nonnegative Grid Factorization", *Journal on Expert Systems with Applications*, 2013 pp. 01–11.

[3] Elena Nenova and Dmitry I. Ignatov, "An FCAbased Boolean Grid Factorization for Collaborative Filtering", *Conference on Formal Concept in information retrieval*, 2012 pp. 57–73.

[4] Alan Ritter, Mausam and Oren Etzioni, "Open Domain Event Extraction from Twitter", *Conference on Knowledge discovery and data mining*, 2012, pp. 1104–1112.

[5] Anish Das Sarma and Alpa Jain, "Dynamic Relationship and Event Discovery",*Conference on Web search and data mining*, 2011, pp. 207–216.

[6] Deng Cai and Jiawei Han, "Graph Regularized Nonnegative Grid Factorization for Data Representation", *Transactions on Pattern analysis and Machine Intelligence*, 2011, Vol. 33, No. 8, pp. 1548–1560.

[7] Abderrahim El Qadi, DrissAboutajdine and YassineEnnouary, „Formal Concept Analysis for Information Retrieval", *Journal of computer science and information security*, 2010, vol.7, No. 2, pp. 119–125.

[8] C. Corley, D. Cook, A. Mikler, and K. Singh, "Text and structural data mining of influenza mentions in web and social media," *Int. J. Environ. Res. Public Health*, , Feb. 2010, vol. 7, pp. 596–615

[9] Fei Wang, "Community discovery using nonnegative grid factorization", *Conference on data mining and data representation*, 2010, pp. 01–29.

[10] Jonas Poelmans, Paul Elzinga and Stijn Viaene, "Formal Concept Analysis in knowledge discovery: a survey",*Conference on conceptual structures*, 2010, pp. 139–153.

[11] Vasumathi, D. and Govardhan, 'Efficient Web usage Mining Based on Formal Concept Analysis', Journal on Theoretical and Applied Information Technology, , 2009, pp. 99–109.

[12] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the Web," *Manag. Sci.*, Sep. 2007, vol. 53, pp. 1375–1388.

[13] P. Soucy and G. W. Mineau, "Beyond TFIDF weighting for text catego-rization in the vector space model," *in Proc. 19th Int. Joint Conf. Artificial Intell., Edinburgh, U.K.*, 2005, pp. 1130–1135.

[14] W. Yih, P. H. Chang, and W. Kim, "Mining online deal forums for hot deals," *in Proc. IEEE/WIC/ACM Int. Conf. Web Intell., Beijing, China*, 2004, pp. 384– 390.

[15] B. Taskar, M. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," *in Proc. Adv. Neural Inform. Process. Syst., Vancouver, B.C. Canada*, 2003.

[16] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "Self-Organizing Map in MATLAB: The SOM Toolbox," *in Proc. Matlab DSP Conf., Espoo*, Finland, 1999, pp. 35–40.