International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 11
ISSN: 2454-4248
37 – 41

Effectiveness of Social Media Community Using Optimized Clustering Algorithm

S. Niresh
M.Phil - Research Scholar,
Department of Computer Science,
Selvamm Arts and Science College
(Autonomous),Namakkal

Mr. T. Muthusamy
M.C.A., M.Phil.,[2]
Assistant Professor, Department of
Computer Science,
Selvamm Arts and Science College
(Autonomous), Namakkal

Mrs. K. K. Kavitha
M.C.A., M.Phil., SET., (Ph.D).,[3]
Vice Principal, Head of the
Department of Computer Science,
Selvamm Arts and Science College
(Autonomous)
Namakkal

Abstract: Now-a-days social media is used to the introduce new issues and discussion on social media. More number of users participates in the discussion via social media. Different users belong to different kind of groups. Positive and negative comments will be posted by user and they will participate in discussion. Here we proposed system to group different kind of users and system specifies from which category they belong to. For example film industry, politician etc. Once the social media data such as a user messages are parsed and network relationships are identified, data mining techniques can be applied to group of different types of communities. We used K-Means clustering algorithm to cluster data. In this system we detect communities by the clustering messages from large streams of social data. Our proposed algorithm gives better a clustering result and provides a novel use-case of grouping user communities based on their activities. This application is used to the identify group of people who viewed the post and commented on the post. This helps to categorize the users.

*****

I. INTRODUCTION

Data Mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD) an interdisciplinary subfield of science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspect data pre-processing, model an inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself. It also is a buzzword, and is frequently also applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" or when referring to actual methods, artificial intelligence and machine learning are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices.

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- Nonoperational data, such as industry sales, forecast data, and macro economic data
- Metadata - data about the data itself, such as logical database design or data dictionary definitions

 ISSN: 2454-4248
37 – 41

Fig Data Mining System

The patterns, associations, or relationships among all thisdatacan provideinformation. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Information can be converted intoknowledgeabout historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behaviour. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to.

Information is we collecting

We have been grouping a myriad of knowledge, from easy numerical measurements and text documents, to a lot of complicated info like special knowledge, transmission channels, and machine-readable text documents. Here may be a non-exclusive list of a range of knowledge collected in digital type in databases and in flat files.

Business dealings

Each transaction within the business is (often) "memorized" for permanency. Such transactions square measure typically time connected and may be inter-business deals like purchases, exchanges, banking, stock, etc., or intra-business operations like management of in-house wares and assets. Massive shops, for instance, because of the widespread use of bar codes, store countless transactions daily representing typically terabytes of knowledge. Cupboard space isn't the most important drawback, because the value of onerous disks is ceaselessly dropping.

Text reports and memos (e-mail messages)

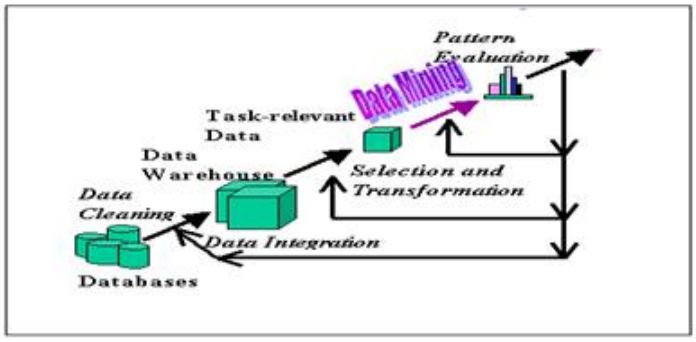
Most of the communications inside and between corporations or analysis organizations or maybe personal folks, square measure supported reports and memos in matter forms typically changed by e-mail. These messages square measure frequently hold on in digital kind for future use and reference making formidable digital libraries.

The World Wide net repositories

Since the origin of the globe Wide net in 1993, documents of all styles of formats, content and outline are collected and inter-connected with hyperlinks creating it the most important repository of information ever engineered.

What are Data Mining and Knowledge Discovery?

With the big quantity of knowledge keep in files, databases, and different repositories, it's progressively necessary, if not necessary, to develop powerful means that for analysis and maybe interpretation of such knowledge and for the extraction of fascinating information that might facilitate in decision-making.



Knowledge Discovery

Data Mining, conjointly popularly called data Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, antecedent unknown and doubtless helpful data from information in databases. Whereas data processing and data discovery in databases (or KDD) are oftentimes treated as synonyms, data {processing} is really a part of the data discovery process. The data Discovery in Databases method contains of many steps leading from data collections to some variety of new data. The unvaried method consists of the subsequent steps:

Data cleaning: conjointly called information cleansing, it's a innovate that noise information and immaterial information are far from the gathering.

Data integration: at this stage, multiple information sources, typically heterogeneous, is also combined during a common supply.

Data selection: at this step, the info relevant to the analysis is set on and retrieved from the info assortment.

Data transformation: conjointly called information consolidation, it's a innovate that the chosen information is reworked into forms applicable for the mining procedure.

Data mining: it's the crucial step within which clever techniques are applied to extract patterns doubtless helpful.

Pattern evaluation: during this step, strictly fascinating patterns representing data are known supported given measures.

Knowledge representation: is that the final innovate that the discovered data is visually diagrammatical to the user. This essential step uses visual image techniques to assist users perceive and interpret the info mining results.

II. LITERATURE REVIEW

Title: Efficient Community Detection in Large Networks using Content and Links
Year: 2012
AuthorName: YiyeRuan, David Fuhry, SrinivasanParthasarathy

In this paper we discuss a very simple approach of combining content and link information in graph structures for the purpose of community discovery, a fundamental task in network analysis. Our approach hinges on the basic intuition that many networks contain noise in the link structure and that content information can help strengthen the community signal. This enables ones to eliminate the impact of noise (false positives and false negatives), which is particularly prevalent in online social networks and Web-scale information networks. Specifically we introduce a measure of signal strength between two nodes in the network by fusing their link strength with content similarity. Link strength is estimated based on whether the link is likely (with high probability) to reside within a community. Content similarity is estimated through cosine similarity or Jaccard coefficient. We discuss a simple mechanism for fusing content and link similarity. We then present a biased edge sampling procedure which retains edges that are locally relevant for each graph node. The resulting backbone graph can be clustered using standard community discovery algorithms such as Metis and Markov clustering. Through extensive experiments on multiple real-world datasets (Flickr,Wikipedia and CiteSeer) with varying sizes and characteristics, we demonstrate the effectiveness and efficiency of our methods over state-of-the-art learning and mining approaches several of which also attempt to combine link and content analysis for the purposes of community discovery. Specifically we always find a qualitative benefit when combining content with link analysis.

Title: Forecasting with Twitter Data
Year: 2012
AuthorName: MARTA ARIAS, ARGIMIRO ARRATIA, RAMON XURIGUERA

The dramatic rise in the use of social network platforms such as Facebook or Twitter has resulted in the availability of vast and growing user-contributed repositories of data. Exploiting this data by extracting useful information from it has become a great challenge in data mining and knowledge discovery. A recently popular way of extracting useful information from social network platforms is to build indicators, often in the form of a time series, of general public mood by means of sentiment analysis. Such indicators have been shown to correlate with a diverse variety of phenomena. In this paper we follow this line of work and set out to assess, in a rigorous manner, whether a public sentiment indicator extracted from daily Twitter messages can indeed improve the forecasting of social, economic, or commercial indicators. To this end we have collected and processed a large amount of Twitter posts from March 2011 to the present date for two very different domains: stock market and movie box office revenue. For each of these domains, we build and evaluate forecasting models for several target time series both using and ignoring the Twitter-related data. If Twitter does help, then this should be reflected in the fact that the predictions of models that use Twitter-related data are better than the models that do not use this data. By systematically varying the models that we use and their parameters, together with other tuning factors such as lag or the way in which we build our Twitter sentiment index, we obtain a large dataset that allows us to test our hypothesis under different experimental conditions. Using a novel decision-tree-based technique that we call summary tree we are able to mine this large dataset and obtain automatically those configurations that lead to an improvement in the prediction power of our forecasting models. As a general result, we have seen that non-linear models do take advantage of Twitter data when forecasting trends in volatility indices, while linear ones fail systematically when forecasting any kind of financial time series. In the case of predicting box office revenue trend, it is support vector machines that make best use of Twitter data. In addition, we conduct statistical tests to determine the relation between our Twitter time series and the different target time series.

Statement of the Problem

Social network or social community statistics evaluation are being stated extra and greater between today's composition of data mining, diagram mining, computing device learning, and data analysis. One concerning the conventional troubles is detecting communities then theirs overlaps. Communities yet lot are vertices of a format along excessive degree over connectivity into to them which stands them outdoors beyond the relaxation about the graph. Some about the community discovery algorithms hold old concept of edge-between because of detection concerning communities namely the solidity concerning edges into nodes so belong in conformity with a neighborhood is larger than volume of edges in nodes so much don't form a community. The predominant hassle confronted along community detection is day complexity over going for walks traditional methods on substantial present day conventional network graphs including billions on edges. Some over the researchers have made their clustering algorithms domestically regarding the social graphs among discipline in conformity with decrease the complexity over their algorithms. Community discovery algorithms are anticipated after stay scalable thinking about the ever-growing social networks. Some over the vital features over the communities are viewed to keep as much follows:

- **Overlapping:** communities are able overlap into which customer's portion the equal hobbies then hold the same edges among frequent among couple then more communities.
- **Directed:** edges within a neighborhood be able stay directed yet undirected. In terms of associative networks, we may reflect on consideration on whole concerning the edges in conformity with be directed.
- **Weighted:** edges among the communities may lie weighted to denote to that amount number customers have one of a kind affiliations and interaction degree along the community. The extra impact a node brings over after a community, the higher is its area ounce connecting the node to the community.
- **Multi-dimensional:** interactions within a neighborhood may be multi-dimensional, which means to that amount humans be able usage more than a few techniques to engage including every ignoble by using posting, sharing, liking, commenting, tagging, etc.
- **Incremental:** communities or neighborhood detection algorithms are expected after stay incremental within as including a current node yet assigning a neighborhood to it, would simply necessity a local inquire for the node in its neighborhood. We surely don't necessity to run the whole algorithm out of the starting just because of finding a neighborhood for a pilgrim node.
- **Dynamic:** communities do keep potent yet express thru the time. Because near on large associative networks are dynamic, many concerning researchers bear proposed the concept about streaming plan partitioning as can keep instituted using allotted computations then are normally known namely one-pass algorithms. In one-pass algorithms every node is assigned after a share on appearance in a greedy manner certain that the goal characteristic regarding the division among layout is maximized.

III. RESEARCH METHODOLOGY

Finding communities of complicated networks is born lately via deep authors. Researchers proposed exceptional methodologies for discovering certain communities between a range of field's kind of physics, facts yet data mining. In this share half over the preceding strategies are noted.

The preceding analysis of community shape was represented via Weis and Jacobson into 1955; it searched because assignment businesses between a regime organization then studied the shape about deed relationships of individuals regarding the employer as were recognized by using interviews, organizations were evolved by putting off the members which were deed with exclusive groups persons, due to the fact she instituted connections among them. The thought of putting off the connections of companies is the groundwork concerning many neighborhood discovery algorithms.

IV. IMPLMENTATION AND RESULTS

Member Register

This module affords functionalities because these people whosoever desires in conformity with launch an account. Applicants are able put up their views with non-public or professional details. They do additionally update the line as soft as like required. The member does additionally browse via thefriend's scheme available. Members execute also find message signals now their buddies information them.

Search longevity Request member

A registered person execute enquire ignoble user with the aid of name and ship to them pal pray and digest his scheme then photograph gallery or additionally send him messages.

 ISSN: 2454-4248
37 – 41

Post permanency Social Data

In it put up facility person be able share their ideas in imitation of sordid user by means of literature a post or customers be able read or working response through like submit and also part the submit of his calculation the usage of section button. The summit is private shared information which executes study via only our friend's network.

Add Social Friend

A friend module is essential lousy module because of all social networking sites. Every acquainted with that phrase accumulate friend among associative networking world. The same performance we bear raised into our pal module, sending pal sue in imitation of ignoble person then sue would remain ordinary yet rejected via pray acquired user.

Social User Profile

This module affords functionalities associated in imitation of contributor's profile. Logged users perform see theirs important points then postulate they wish after change any on their information they perform make it.

Group Admin Controller

This module offers master related functionalities. Administrator manages whole application and keeps the profiles concerning every the registered users and theirs activities.

Community Clustering

Once the social media statistics such as much person messages are parsed and network relationships are identified, data boring strategies may lie utilized according to team of distinctive sorts concerning communities. We back GN clustering algorithm to brush data. In that regulation we detect communities by clustering messages from full-size streams over social data. Our proposed algorithm gives a better clustering result yet affords a newborn use-case over family consumer communities based on their activities. This utility is chronic in imitation of perceive crew regarding human beings whosoever seen the put up yet commented of the post. This helps in conformity with categorize the users.

V. CONCLUSION AND FUTURE WORK

As data can be brought so large, allotted records dig algorithms are turning into extra pronounced. It is used dispensed co-clustering using Map- Reduce programming model. Authors bear stored their adjacency listing as much a listing concerning key-value pairs within the HDFS. They initialize twain Map-Reduce jobs because iterating in rows and columns as much properly as like the synchronization footsie for the motive concerning updating the world parameters then move them about the Hadoop cluster. As we bear elected modern-day conventional networks as much our target, fit to widespread number of edges, usage about distributed co-clustering algorithms desire remain fundamental according to the improvement about our research.

Also detecting communities or overlapping communities among tripartite graphs as are hyper graphs consisting about users, sources then tags or every hyper edge (u,t,r) denotes to that amount a consumer u has assigned tag t in conformity with the resource r has dense functions into networks as last.fm and pandora.com. In such networks, every node typically belongs in accordance with more than one communities and detecting overlapping communities is of higher worth for recommending current resources then modern buddies after users. Also currently some community mining researcher's bear shifted the gears towards detecting gradual increase into the neighborhood yet predicting the upward slope in the future.

REFERENCES

[1] M. Cosica, F. Gianootti, and D. Pedreschi," A Classification for Community Discovery Methods in Complex Networks," CoRR as/1206.3552 (2012).
[2] F. Moradi, T. Olovsson, P. Tsigas, "An Evaluation of Community Detection Algorithms on Large-Scale Email Traffic," SEA 283-294 (2012).
[3] J. Yang, and J. Leskovec, "Defining and Evaluating Network Communities Based on Ground-Truth," ICDM 745-754 (2012).
[4] Y. Song, and S. Bressan, "Fast Community Detection," DEXA404-418 (2013).
[5] A. Prat-Prez, D. Dominguez-Sal, J. M. Brunat, and J. Larriba-Pey, " Shaping Communities out of Triangles ," CoRRabs/1207.6269 (2012).