Finding Burst Content In Online Stream
Via URL Based Detection

KOTIPALLI SUMA
Department of Computer Science and Engineering
M.Tech Student, Sri Venkateswara College Of
Engineering and Technology

D. RAMYA
Department of Computer Science and Engineering
Assistant professor, Sri Venkateswara College of
Engineering and Technology

Abstract: Detection of emerging topics is now receiving renewed interest motivated by the rapid growth of social networks. Conventional-term-frequency-based approaches may not be appropriate in this context, because the information exchanged in social-network posts include not only text but also images, URLs, and videos. We focus on emergence of topics signaled by social aspects of these networks. Specifically, we focus on mentions of user links between users that are generated dynamically (intentionally or unintentionally) through replies, mentions, and retweets. We propose a probability model of the mentioning behavior of a social network user, and propose to detect the emergence of a new topic from the anomalies measured through the model. Aggregating anomaly scores from hundreds of users, we show that we can detect emerging topics only based on the reply/mention relationships in social-network posts. We demonstrate our technique in several real data sets we gathered from Twitter. The experiments show that the proposed mention-anomaly-based approaches can detect new topics at least as early as text-anomaly-based approaches, and in some cases much earlier when the topic is poorly identified by the textual contents in posts.

I. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

What is Data Mining?

![Structure of Data Mining](image)

How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

1) Extract, transform, and load transaction data onto the data warehouse system.

2) Store and manage the data in a multidimensional database system.

3) Provide data access to business analysts and information technology professionals.

4) Analyze the data by application software.
5) Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where $k=1$). Sometimes called the $k$-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

**Characteristics of Data Mining:**

- **Large quantities of data**: The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.

- **Noisy, incomplete data**: Imprecise data is the characteristic of all data collection.

- **Complex data structure**: Conventional statistical analysis not possible.

- **Heterogeneous data stored in legacy systems**

**Benefits of Data Mining:**

1) It’s one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them

2) An analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customers

3) An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)

4) Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors

5) Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek

**Advantages of Data Mining:**

1. **Marketing / Retail**:
   Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign...etc. Through the results, marketers will have an appropriate approach to sell profitable products to targeted customers.

   Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

2. **Finance / Banking**
   Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer’s data,
the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card’s owner.

3. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example, semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

4. Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

5. Law enforcement:

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

6. Researchers:

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

ORGANIZATION PROFILE

COMPANY PROFILE:

Founded in 2009, JP iNFOtCH located at Puducherry, has a rich background in developing academic student projects, especially in solving latest IEEE Papers, Software Development and continues its entire attention on achieving transcending excellence in the Development and Maintenance of Software Projects and Products in Many Areas.

The indispensable factors, which give the competitive advantages over others in the market, may be slated as:

- Performance
- Pioneering efforts
- Client satisfaction
- Innovative concepts
- Constant Evaluations
- Improvisation
- Cost Effectiveness

ABOUT THE PEOPLE:

As a team we have the clear vision and realize it too. As a statistical evaluation, the team has more than 40,000 hours of expertise in providing real-time solutions in the fields of Android Mobile Apps Development, Networking, Web Designing, Secure Computing, Mobile Computing, Cloud Computing, Image Processing And Implementation, Networking With OMNET++, Simulator, client Server Technologies in Java,J2EE,J2ME,EJB, ANDROID, DOTNET (ASP.NET, VB.NET, C#.NET), MATLAB, NS2, SIMULINK, EMBEDDED, POWER ELECTRONICS, VB & VC++, Oracle and operating system concepts with LINUX.

II. EXISTING SYSTEM:

- A new (emerging) topic is something people feel like discussing, commenting, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words.

DISADVANTAGES OF EXISTING SYSTEM:

A term-frequency-based approach could suffer from the ambiguity caused by synonyms or homonyms. It may also require complicated preprocessing (e.g., segmentation) depending on the target language. Moreover, it cannot be applied when the contents of the messages are mostly nontextual information. On the other hand, the “words” formed by mentions are unique, require little preprocessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents.

III. PROPOSED SYSTEM:

- In this paper, we have proposed a new approach to detect the emergence of topics in a social network stream.
- The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behaviour of users instead of the textual contents.
- We have proposed a probability model that captures both the number of mentions per post and the frequency of mentionee.

ADVANTAGES OF PROPOSED SYSTEM:

- The proposed method does not rely on the textual contents of social network posts, it is robust to rephrasing and it can be applied to the case where topics are concerned with information other than
texts, such as images, video, audio, and so on.

- The proposed link-anomaly-based methods performed even better than the keyword-based methods on “NASA” and “BBC” data sets.

MODULES:
1. Training
2. Identify individual Anomaly Score
3. Aggregate
4. Change Point Analysis and DTO
5. Burst Detection

SYSTEM REQUIREMENTS:

HARDWARE REQUIREMENTS:
- System: Pentium IV 2.4 GHz.
- Hard Disk: 40 GB.
- Floppy Drive: 1.44 Mb.
- Monitor: 15 VGA Colour.
- Mouse: Logitech.
- Ram: 512 Mb.

SOFTWARE REQUIREMENTS:
- Operating system: Windows XP/7.
- Coding Language: JAVA/J2EE
- IDE: Netbeans 7.4
- Database: MYSQL

IV. TECHNOLOGIES USED

4.1 Introduction to Java:
Java has been around since 1991, developed by a small team of Sun Microsystems developers in a project originally called the Green project. The intent of the project was to develop a platform-independent software technology that would be used in the consumer electronics industry. The language that the team created was originally called Oak.

The first implementation of Oak was in a PDA-type device called Star Seven (*7) that consisted of the Oak language, an operating system called GreenOS, a user interface, and hardware. The name *7 was derived from the telephone sequence that was used in the team's office and that was dialed in order to answer any ringing telephone from any other phone in the office.

Java technology is both a programming language and a platform.

The Java Platform:
A platform is the hardware or software environment in which a program runs. We’ve already mentioned some of the most popular platforms like Windows 2000, Linux, Solaris, and Mac OS. Most platforms can be described as a combination of the operating system and hardware. The Java platform differs from most other platforms in that it’s a software-only platform that runs on top of other hardware-based platforms.

The Java platform has two components:
- The java virtual machine (Java VM)
- The java application programming interface (Java API)

You’ve already been introduced to the Java VM. It’s the base for the Java platform and is ported onto various hardware-based platforms.

The Java API is a large collection of ready-made software components that provide many useful capabilities, such as graphical user interface (GUI) widgets. The Java API is grouped into libraries of related classes and interfaces; these libraries are known as packages. The next section, What Can Java Technology Do?, highlights what functionality some of the packages in the Java API provide.

The following figure depicts a program that’s running on the Java platform. As the figure shows, the Java API and the virtual machine insulate the program from the hardware.

Figure 4.2: The Java Platform

Working Of Java:
For those who are new to object-oriented programming, the concept of a class will be new to you. Simplistically, a class is the definition for a segment of code that can contain both data and functions. When the interpreter executes a class, it looks for a particular method by the name of main, which will sound familiar to C programmers. The main method is passed as a parameter an array of strings (similar to the argv[] of C), and is declared as a static method.

To output text from the program, execute the println method of System.out, which is java’s output stream. UNIX users will appreciate the theory behind such a stream, as it is actually standard output. For those who are instead used to the Wintel platform, it will write the string passed to it to the user’s program.
FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

• ECONOMICAL FEASIBILITY
• TECHNICAL FEASIBILITY
• SOCIAL FEASIBILITY

Functional and Non-Functional Requirements:

1. Functional Requirements:
   a. Inputs:
      Browsing and uploading of files.
   b. Processing:
     Load server: Stores all files
     Slip server cluster:
     • Browses the file
     • Selects the path
     • Download the file

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy.

Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system’s relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find

---

**Figure 4.3 knowledge discovery process**
the system can use easily and effectively. When analysis design computer output, they should identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

VII. SYSTEM DESIGN

SYSTEM ARCHITECTURE:

BLOCK DIAGRAM

UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Metamodel and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing object oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message
Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

**ACTIVITY DIAGRAM:**

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

**COLLABORATIVE DIAGRAM:**

Testing is a process of executing a program with the intent of finding an error. A good test case is one that has a high probability of finding an as-yet undiscovered error. A successful test is one that uncovers an as-yet undiscovered error. System testing is the stage of implementation, which is aimed at ensuring that the system works accurately and efficiently as expected before live operation commences. It verifies that the whole set of programs hang together. System testing requires a test consists of several key activities and steps for run program, string, system and is important in adopting a successful new system. This is the last chance to detect and correct errors before the system is installed for user acceptance testing.

**VIII. CONCLUSION**

In this paper, we have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. We have proposed a probability model that captures both the number of mentions per post and the frequency of mentionee. We have combined the proposed mention model with the SDNML change-point detection algorithm [3] and Kleinberg’s burst-detection model [2] to pinpoint the emergence of a topic. Since the proposed method does not rely on the textual contents of social network posts, it is robust to rephrasing and it can be applied to the case where topics are concerned with information other than texts, such as images, video, audio, and so on. We have applied the proposed approach to four real data sets that we have collected from Twitter. The four data sets included a wide-spread discussion about a controversial topic (“Job hunting” data set), a quick propagation of news about a video leaked on Youtube (“Youtube” data set), a rumor about the upcoming press conference by NASA (“NASA” data set), and an angry response to a foreign TV show (“BBC” data set). In all the data sets, our proposed approach showed promising performance. In three out of four data sets, the detection by the proposed link-anomaly-based methods was earlier than the text-anomaly-based counterparts. Furthermore, for “NASA” and “BBC” data sets, in which the keyword that defines the topic is more ambiguous than the first two data sets, the proposed link-anomaly-based approaches have detected the emergence of the topics even earlier than the keyword-based approaches that use hand-chosen keywords. All the analysis presented in this paper was conducted offline, but the framework itself can be applied online. We are planning to scale up the proposed approach to handle social streams in real time. It would also be interesting to combine the proposed link-anomaly model with text-based approaches, because the proposed link-anomaly model does not immediately tell what the anomaly is. Combination of the word-based approach with
the link-anomaly model would benefit both from the performance of the mention model and the intuitiveness of the word-based approach.

**IX. REFERENCES**


**AUTHOR'S PROFILE**

**KOTIPALLI SUMA**, Department of Computer Science and Engineering
M.Tech Student, Sri Venkateswara College Of Engineering and Technology

**D. RAMYA**, Department of Computer Science and Engineering
Assistant professor, Sri Venkateswara College of Engineering and Technology