TDPA: Trend Detection and Predictive Analytics

M. Sakthi ganesh[1], CH.Pradeep Reddy[2], N.Manikandan[3], DR.P.Venkata krishna[4]

1. Assistant Professor , School of Information Technology & Engineering (SITE), VIT University, Vellore,
2. Assistant Professor, School of Information Technology & Engineering (SITE), VIT University ,Vellore,
3. Assistant Professor(senior) , School of Information Technology& Engineering (SITE), VIT University, Vellore,
4. Professor, School of Computing sciences & Engineering,VIT University,Vellore

Abstract - Text mining is the process of exploratory text analysis either by automatic or semi-automatic means that helps finding previously unknown information. Text mining is a highly interdisciplinary research area, bringing together research insights from the fields of data mining, natural language processing, machine learning, and information retrieval. The amount of textual data available is too huge to be managed manually. An automatic system is needed to analyze and interpret the text. Some of the systems are semi automatic requiring user input to begin processing others are fully automatic producing output from the input corpus without guidance. The review literatures on trend detection indicates that much progress has been made toward automating the process of detecting emerging trends but there is room for improvement. In this work, we propose a Trend Detection and Predictive Analytics (TDPA) using Living Analytics to detect emerging trends from live data to cater the needs of various users irrespective of their domain. The system needs to serve as general purpose software that will help the users to identify and visualize current happenings pertaining to any domain in an efficient and user friendly way. The paper also aims at forecasting the future of the trends obtained in helping the users to look forward and make quick decisions.

Keywords:- mining, detect trend, predictive analytics.

I. Introduction

Text mining is challenging mainly due to the characteristics of text. Text is not well structured, and text data could be noisy. It has high dimensionality. There is dependency in the text, that is, relevant information is a complex conjunction of words/phrases. Moreover, the text being analyzed will have word and semantic ambiguity. Figure 1 shows the major phases involved in text mining and each phase is being discussed below. Raw data which is unstructured in nature and of varied form is inputted to the process and undergoes the following phases to obtain patterns that are meaningful and useful.

Predictive analytics encompasses a variety of techniques from statistics and data mining that process current and historical data in order to make “predictions” about future events. Such predictions rarely take the form of absolute statements, and are more likely to be expressed as values that correspond to the odds of a particular event or behavior taking place in the future. In business, the models often process historical and transactional data to identify the risk or opportunity associated with a specific customer or transaction. These analyses weigh the relationship between many data elements to isolate each customer’s risk or potential, which guides the action on that customer.

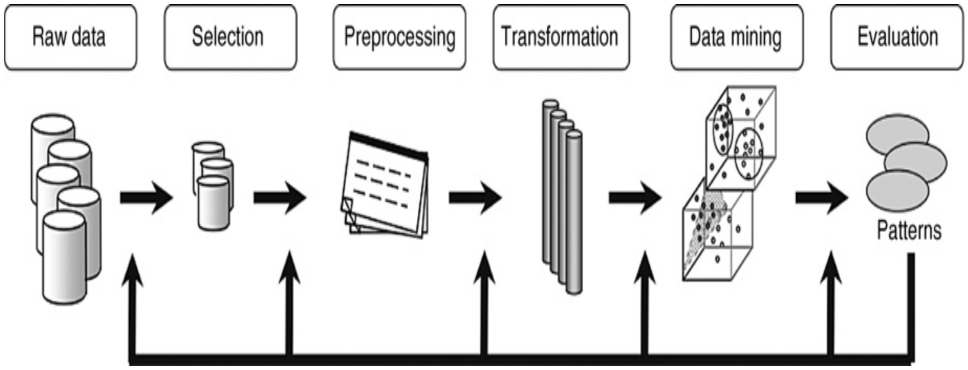


Figure 1 Phases of Text Mining

Predictive analytics is widely used in making customer decisions. One of the most well-known applications is credit scoring, which is used throughout financial services. Scoring models process a customer's credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time. Predictive analytics are also used in insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and other fields.

Existing Systems that Detect Trends

An emerging trend is a topic area that is growing in interest and utility over time. For example, Extensible Markup Language (XML) emerged as a trend in the mid 1990s. A detailed description of several semi automatic and fully automatic ETD systems used for research purposes or educational purposes is as follows.

TOA (Technology Opportunities Analysis) - TOA is a semi automatic trend detection system for technology opportunities analysis. Input to the system would be abstracts from technical database such as INSPEC, COMPENDEX, US patents. Potential keywords from the abstracts are extracted manually by domain experts. These keywords are then combined into queries using appropriate Boolean operators to generate comprehensive and accurate searches. The queries are then input to the Technology Opportunities Analysis Knowbot (TOAK), a custom software package also referred to as TOAS Technology Opportunities Analysis System. TOAK extracts the relevant documents abstracts and provides analysis of the data by using information such as word counts, date information, word co occurrence information, citation information and publication information to track activity in a subject area. Trend detection is left to the user in this semi-automatic method.

THEME RIVER - Theme River is yet another trend detection tool that summarizes the main topics in a corpus and presents a summary of the importance of each topic via a graphical user interface. The topical changes over time are shown as a river of information. The river is made up of multiple streams. Each stream represents a topic and each topic is represented by a color and maintains its place in the river relative to other topics. Like TOA and Time Mines Theme River does not presume to indicate which topics are emergent. The visualization is intended to provide the user with information about the corpus.

PATENT MINER - The Patent Miner system was developed to discover trends in patent data using a dynamically generated SQL query based upon selection criteria input by the user. The system is connected to an IBM DB2 database containing all granted United States (US) patents. There are two major components to the system phrase identification using sequential pattern mining and trend detection using shape queries.

Several semi automatic and fully automatic ETD systems providing detailed information relating to linguistic and statistical features training and test set generation learning algorithms has been discussed above. It indicates that much progress has been made towards automating the process of detecting emerging trends but there remains room for improvement. All of the systems rely on human domain expertise to separate emerging trends from noise in the system. In addition few systems whether research or commercial in nature have employed formal evaluation metrics and methodologies to determine effectiveness. The development and use of metrics for evaluation of ETD systems is critical.

Existing Systems for Predictive Analytics

Predictive Analytics is a branch of business intelligence category, uses data mining and statistics to make predictions on future happenings. The predictions tell you what are the odds that a certain event will be taking place or not, under what circumstances, or following trends. Description of few open source Predictive Analytics is discussed below.

RAPID MINER - Rapid Miner (formerly YALE (Yet Another Learning Environment)) is an environment for machine learning and data mining experiments. It allows experiments to be made up of a large number of arbitrarily nestable operators, described in XML files which are created with Rapid Miner's graphical user interface. Rapid Miner is used for both research and real-world data mining tasks. Rapid Miner provides a GUI to design an analytical pipeline. The GUI generates an XML (eXtensible Markup Language) file that defines the analytical processes the user wishes to apply to the data. This file is then read by Rapid Miner to run the analyses automatically.

WEKA - Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Overall Design

On determining the purpose and specification of the project, the design of the project is accomplished to develop plan for the obtained solution. Considering various aspects of the software such as compatibility, extensibility, fault - tolerance, maintainability, modularity, reusability and usability the following design has been constructed.

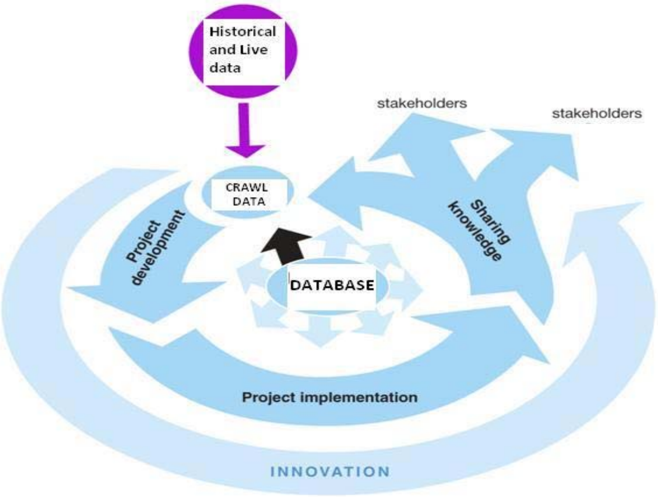


Figure 2 Overall Design

Figure 2 shows the overall design (Higher-level) of the system to be implemented. Figure 3 shows various phases (Functional level design) that have been analyzed and identified. Trend Detection module consists of a Back-End and a Front-End layer (Figure 4). From Back-End the Carrot^2 Clustering engine pulls data from various live sources like W3, Wiki, News, and Blogs via respective APIs according to the keyword inputted by the end user using the UI. Carrot^2 can currently sample upto150 document links. Every document is clustered using “Lingo” clustering algorithm. Then each cluster is being processed individually to obtain the most frequently used terms within each cluster. Threshold limit for most frequent words is set to be 10% from the term that is used highest. Similarly each cluster is being processed and finally obtained keywords are classified as it belongs to the trend category or not. Further analysis on the obtained trends is performed and top rated trends are sent to the UI where the user visualizes various trends obtained.

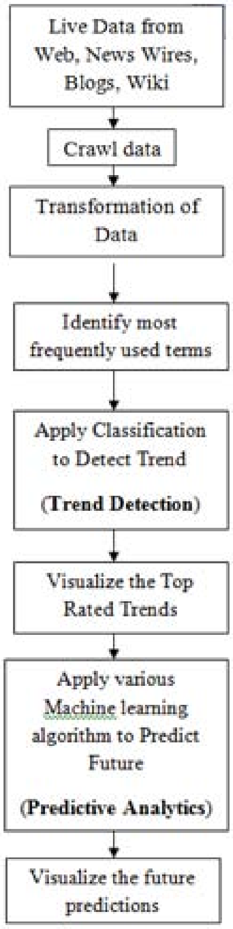

Live Data from Web, News Wires, Blogs, Wiki
Crawl data
Transformation of Data
Identify most frequently used terms
Apply Classification to Detect Trend
(Trend Detection)
Visualize the Top Rated Trends
Apply various Machine learning algorithm to Predict Future
(Predictive Analytics)
Visualize the future predictions

Figure 3 phases of the project

Figure 4 shows the various steps involved in detecting the trend for the keyword given. The proposed system is expected to perform predictive analytics (Figure 5) from a large input corpus which is predominantly textual in nature and hence it could also be stated as text analytics. Like data mining, text analytics is an iterative process, and is most effective when it follows a proven methodology. This maximizes analyst productivity, supports comparability of results, allows findings from one analysis to be used to inform or guide others, and facilitates data-driven decision making. As with data mining, the two main steps in text analytics are data preparation and data understanding.

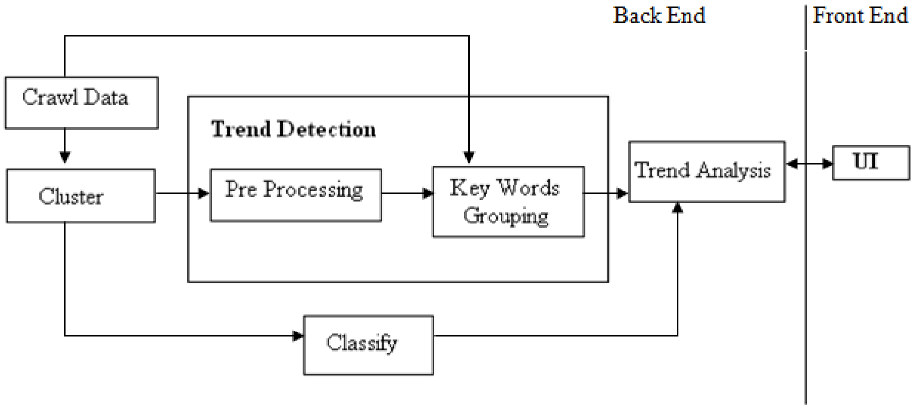


Figure 4 Trend detection Architecture

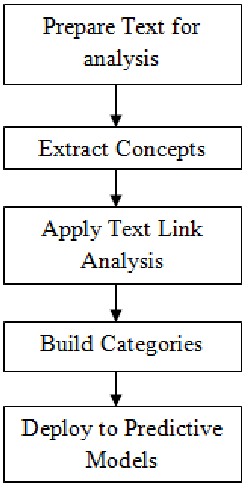


Figure 5 Predictive Analytics Process

Figure 5 shows various phases and concepts involved in Predictive Analytics Module.

II. Implementation and Results

The proposed algorithm has its base in text mining where live data from various sources like web, blogs, new wires etc., needs to be crawled to get only the relevant pages. Once the relevant pages of text are obtained, those pages should be mined for detection of topics. Various text mining preprocesses like stop word removal, stemming, POS tagging, disambiguation are to be performed to clean the text obtained. Later text mining techniques like clustering, classification etc., and various other statistical methods are applied to categorize the text and obtain relevant topics. Trend detection algorithms are to be applied to the detected topics to identify the current trend in the topics obtained. Various predictive analytics methodologies need to be applied to the obtained trends to predict its future. The final output i.e. the predicted future trends needs to be presented to the user of the system via various visualization techniques. The proposed algorithm has five major steps:

- Crawling of Live data from various Websites
- Topic detection
- Trend Detection
- Visualization of the trends and their expected future

Crawling Data

The system under development is not domain specific and requires data that is both historic as well as up to date i.e. Live Data. Owing to the above constraints data cannot be stored in Databases for use. Instead input data has to flow into the system from the enormous digital resources that are readily available from sources such as World Wide Web. Since Trend detection requires information on current situation, websites that specifically hold data on current news like news wires are also to be chosen. Such kind of data has to be crawled into the system using Web Crawlers. A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses. Owing to resource constraints and goal of the project, crawling data from web which by itself is a huge project cannot be implemented in the provided short span. Thus this issue was analyzed and use of Carrot^2 tool for crawling data from web was employed.

Carrot^2 is being used for crawling of data and is not a search engine itself; it does not have a crawler and indexer. It calls the respective Website's API which is inbuilt and uses the crawler that is supported by that website. Carrot^2 is open source and the challenge was to integrate it with Eclipse environment for further processing.

Topic Detection

TDT (Topic Detection and Tracking) study is intended to explore techniques for detecting the appearance of new topics and for tracking the reappearance and evolution of them. TDT study assumes multiple sources of information, for example various newswires and various news broadcast programs. The information flowing from each source is assumed to be divided into a sequence of stories, which may provide information on one or more events. The general task is to identify the events being discussed in these stories, in terms of the stories that discuss them.

Trend Detection

The system performs trend detection in two steps and analyzes trends in a third step. First, it groups topics into clusters based on their co-occurrences. Then it identifies 'frequent' keywords, i.e. keywords that suddenly appear in the cluster at an unusually high rate. In other words, a trend is identified as a set of frequently occurring keywords. After a trend is identified, the system proposes to extract additional information from the documents that belong to the topic, aiming to discover interesting aspects of it. The system initially aims at selecting target terms from the inputted corpus. First the statements need to be tokenized. Tokenization is the task of chopping the statements into pieces, called *tokens*, perhaps at the same time throwing away certain characters, such as punctuation. The presence of white space surrounding a single character or a group of characters defines an explicit token (a word).

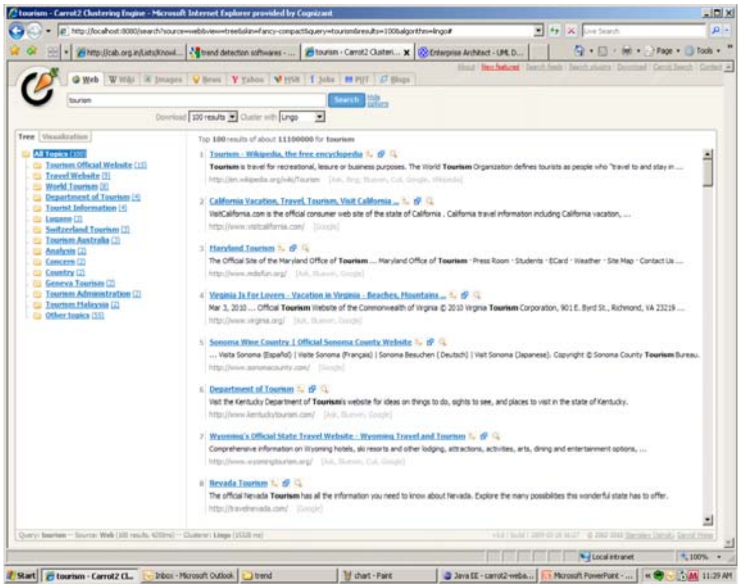
Input: Friends, Romans, Countrymen, lend me your ears.
Output:

Friends	Romans	Countrymen	lend	me	your	ears

Once the corpus is tokenized various stop words among the tokens needs to be eliminated. Common words like *a*, *the*, *which*, *where etc.*, are removed to increase performance. The frequency count of the words obtained is calculated and the maximum frequency count is computed. With the maximum frequency count a threshold value is fixed. The algorithm retains those terms that have a distribution larger than this threshold fixed. Now each of the keyword obtained has to undergo a series of stages to know if it is the trend or not.

Deploying results to predictive model - Deployment of text analytics results to predictive models is the step that will link text analytics to decision making.

The above mentioned steps and processes of Predictive analytics is the proposed approach that has to be implemented into the system and visualized.


tourism - Carrot2 Clustering Engine - Microsoft Internet Explorer provided by Cognizant
http://localhost:8080/search?source=web&view=tree&skin=fancy-compact&query=tourism&results=100&algorithm=lingo#
Live Search
http://cab.org.in/Lists/Knowl...
trend detection softwares - ...
tourism - Carrot2 Clusteri...
Enterprise Architect - UML D...
Page
Tools
Web
Wiki
Images
News
Yahoo
MSN
Jobs
PUT
Blogs
Search
Hide options
Tree
Visualization
Local intranet
100%
Start
tourism - Carrot2 Cl...
Inbox - Microsoft Outlook
trend
chart - Paint
Java EE - carrot2-weba...
Microsoft PowerPoint - ...
11:29 AM

tourism

Download: 100 results Cluster with: Lingo

All Topics (100)
- Tourism Official Website (15)
- Travel Website (9)
- World Tourism (8)
- Department of Tourism (4)
- Tourist Information (4)
- Lugano (3)
- Switzerland Tourism (3)
- Tourism Australia (3)
- Analysis (2)
- Concern (2)
- Country (2)
- Geneva Tourism (2)
- Tourism Administration (2)
- Tourism Malaysia (2)
- Other topics (55)

Top 100 results of about 11100000 for tourism

1 Tourism - Wikipedia, the free encyclopedia
Tourism is travel for recreational, leisure or business purposes. The World Tourism Organization defines tourists as people who "travel to and stay in ...
http://en.wikipedia.org/wiki/Tourism [Ask, Bing, Bluevin, Cuil, Google, Wikipedia]

2 California Vacation, Travel, Tourism, Visit California ...
VisitCalifornia.com is the official consumer web site of the state of California . California travel information including California vacation, ...
http://www.visitcalifornia.com/ [Google]

3 Maryland Tourism
The Official Site of the Maryland Office of Tourism ... Maryland Office of Tourism · Press Room · Students · ECard · Weather · Site Map · Contact Us ...
http://www.mdisfun.org/ [Ask, Bluevin, Google]

4 Virginia Is For Lovers - Vacation in Virginia - Beaches, Mountains ...
Mar 3, 2010 ... Offical Tourism Website of the Commonwealth of Virginia © 2010 Virginia Tourism Corporation, 901 E. Byrd St., Richmond, VA 23219 ...
http://www.virginia.org/ [Ask, Bluevin, Google]

5 Sonoma Wine Country | Official Sonoma County Website
... Visita Sonoma (Español) | Visite Sonoma (Français) | Sonoma Besuchen (Deutsch) | Visit Sonoma (Japanese). Copyright © Sonoma County Tourism Bureau.
http://www.sonomacounty.com/ [Google]

6 Department of Tourism
Visit the Kentucky Department of Tourism's website for ideas on things to do, sights to see, and places to visit in the state of Kentucky.
http://www.kentuckytourism.com/ [Ask, Bluevin, Google]

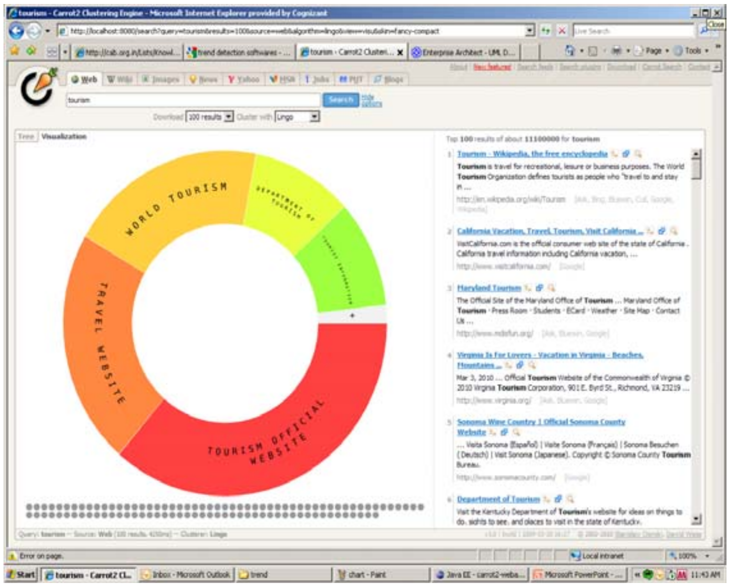
7 Wyoming's Official State Travel Website - Wyoming Travel and Tourism
Comprehensive information on Wyoming hotels, ski resorts and other lodging, attractions, activities, arts, dining and entertainment options, ...
http://www.wyomingtourism.org/ [Ask, Bluevin, Cuil, Google]

8 Nevada Tourism
The official Nevada Tourism has all the information you need to know about Nevada. Explore the many possibilites this wonderful state has to offer.
http://travelnevada.com/ [Google]

Query: tourism -- Source: Web (100 results, 420ms) -- Clusterer: Lingo (1528 ms)

Figure 6 Clustered document in tree view

Figure 6 shows snapshot of the system with keyword “tourism” inputted by the user. The display screen shows the clusters on the left pane and the documents in the right pane.


tourism - Carrot2 Clustering Engine - Microsoft Internet Explorer provided by Cognizant
http://localhost:8080/search?query=tourism&results=100&source=web&algorithm=lingo&view=visu&skin=fancy-compact
Live Search
Close
http://cab.org.in/Lists/Knowl...
trend detection softwares - ...
tourism - Carrot2 Clusteri...
Enterprise Architect - UML D...
Page
Tools
Web
Wiki
Images
News
Yahoo
MSN
Jobs
PUT
Blogs
tourism
Search
Download
100 results
Cluster with
Lingo
Tree
Visualization
WORLD TOURISM
DEPARTMENT TOURISM OF
TRAVEL WEBSITE
TOURISM OFFICIAL WEBSITE
Error on page.
Local intranet
100%
Start
tourism - Carrot2 Cl...
Inbox - Microsoft Outlook
trend
chart - Paint
Java EE - carrot2-weba...
Microsoft PowerPoint - ...
11:43 AM

Top 100 results of about 11100000 for tourism

1 Tourism - Wikipedia, the free encyclopedia
Tourism is travel for recreational, leisure or business purposes. The World Tourism Organization defines tourists as people who "travel to and stay in ...
http://en.wikipedia.org/wiki/Tourism

2 California Vacation, Travel, Tourism, Visit California ...
VisitCalifornia.com is the official consumer web site of the state of California . California travel information including California vacation, ...
http://www.visitcalifornia.com/ [Google]

3 Maryland Tourism
The Official Site of the Maryland Office of Tourism ... Maryland Office of Tourism · Press Room · Students · ECard · Weather · Site Map · Contact Us ...
http://www.mdisfun.org/

4 Virginia Is For Lovers - Vacation in Virginia - Beaches, Mountains ...
Mar 3, 2010 ... Official Tourism Website of the Commonwealth of Virginia © 2010 Virginia Tourism Corporation, 901 E. Byrd St., Richmond, VA 23219 ...
http://www.virginia.org/

5 Sonoma Wine Country 1 Official Sonoma County Website
... Visita Sonoma (Español) | Visite Sonoma (Français) | Sonoma Besuchen (Deutsch) | Visit Sonoma (Japanese). Copyright © Sonoma County Tourism Bureau.
http://www.sonomacounty.com/ [Google]

6 Department of Tourism
Visit the Kentucky Department of Tourism's website for ideas on things to do, sights to see, and places to visit in the state of Kentucky.

Figure 7 Cluster visualization

Figure 7 shows snapshot of the system with keyword “tourism” inputted by the user. The display screen shows the clusters in the left pane and the clicked document in the right pane. Thus the chapter gives the current state of implementation and also discusses on the proposed approaches of the modules to be implemented.

V. Conclusions

This paper analysis on how trends could be detected from live data that is not domain specific has been accomplished fully. A detailed level of system design is formulated from the analysis and research undergone. Many open source tools that will help in the implementation are also given a glimpse to. The approach towards predictive analytics and its visualization are under progress. Trend detection part is successfully implemented and further enhancement would be identifying the association between various trends identified. The process of emerging trend detection can be made iterative. As of now Carrot^2 tool takes only an input of 150 documents at its maximum, thus steps can be taken to scale Carrot^2 for better precision in the trends obtained. Predictive Analytics has been analyzed and designed completely. Implementation of the same requires detailed study and research of the domain as it is an emerging and promising field. On successful implementation, it would serve as a great means of decision making to all the stakeholders involved with the system. Efficient visualization of the same has to be implemented for better presentation and understanding.

References

[1]. Debbie Mayville, “Using predictive Analytics to uncover root causes and solve problems Vs. Treatment symptoms”, August2, 2006.
[2]. A. Berger, S. Della Pietra, and V. Della Pietra, “A maximum entropy approach to natural language processing”, Computational Linguistics, 22 (1) : 39 – 71, 1996.
[3]. J. M. Ponte, and W.B. Croft, “Text Segmentation by Topic”, Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pp. 120–129, 1997.
[4]. Soma Roy, David Gevry, and William M. Pottenger, “Methodologies for Trend Detection in Textual Data Mining”, 2000.
[5]. Kleinberg J. Bursty, “Hierarchical Structure in Streams”, ISIGKDD'02, Edmonton, Alberta, Canada, 2002.
[6]. Stanislaw Osinski, and Dawid Weiss, “Clustering Search Results with Carrot”, January 2007.
[7]. Michael Mathioudakis, and Nick Koudas, “Twitter Monitor: Trend Detection over the Twitter Stream”, 2009.
[8]. Charles Nyce, “Predictive Analytics White Paper”, 2007.
[9]. April Kontostathis, Lars E. Holzman, and William M. Pottenger, “Use of Term Clusters for Emerging Trend Detection”.
[10]. http://openpdf.com/ebook/trend-detection-pdf-2 .html
[11]. http://maroo.cs.umass.edu/pub/web/getpdf.php?id=14
[12]. http://abbottanalytics.blogspot.com/
[13]. http://acca-pakistan.com/2010/03/using-predictive-analytics-within-business-intelligence
[14]. http://en.wikipedia.org/wiki/web-crawler
[15]. http://rapid-i.com/content/view/26/84/