News Source Evaluation and Visualization System

Abdullah Al-Barakati
Information System Department,
Faculty of Computing and
Information Technology
King Abdulaziz University,
Jeddah, Saudi Arabia

ABSTRACT

The advent of Web 2.0 and modern internet technologies was associated with a revolution in online contents especially those that are news-oriented. The proliferation of an array of online mediums for news dissemination and aggregation led to the availability of a constant flow of news contents to end-users. Moreover, the advent of Social Media (SM) platforms meant that the Web was transformed to the biggest live news platform on the planet. In an Era when online users are exposed to an overwhelming stream of news contents, there is an increasing need for tools to empower them to quickly glean and digest the information that they are looking for. Users also need effective tools for news intelligence and prediction to keep abreast of their interests. This paper proposes an interactive online system that will enable users to have an analytical view on the news feeds that are related to their interests. The Visual News Screener (VNS) will aim to surpass the traditional news aggregation systems by its ability to evaluate the effectiveness in which a given news source covers a certain news topic or issue. VNS will have the flexibility of analyzing a plethora of news sources and visually summarize the aggregated data within a customized dashboard (the News Screener). The news contents that VNS will analyze will include RSS feeds, SM feeds, crawled online news sources and articles. The visualization process will depend on the actual context that the user is interested in. Furthermore, the history, current status and potential development patterns of the monitored news issue(s) will also be analyzed and visualized.

Keywords

news analysis, news visualization, online news, news intelligence, source efficiency

1. INTRODUCTION

It can be argued that the World Wide Web (WWW) is one of the richest, most varied and most accessible public information resources [1]. The World Wide Web (referred to from now on as the Web) has profoundly affected the way that news is being produced and disseminated. Online news contents have become a significant part of the social, economic, and cultural life in many societies nowadays [2]. Moreover, news-related contents are considered to be among the most popular types of contents for online users [3]. A number of factors have contributed to the growing influence and popularity of online news contents [4]. These popularity factors primarily include the decline in newspaper readership and advertising revenues [5]. Another popularity factor is the ease of access and convenience the online news platforms provide their users with.

The decline in newspaper readership led news vendors to utilize online news platforms. These platforms are considered to be efficient and cost-effective publishing mediums as opposed to traditional paper-based news publishing [6]. A recent example is the Independent newspaper which is the first British newspaper title to move to a digital-only future [7]. Another contributing factor is the relative ease at which online news contents can be accessed. The multiple channels through which online news contents can be viewed (PCs, laptops, tablets, smart phones, etc) [8] further fueled the strong public preference towards accessing news online.

The rapid growth in online news contents was accompanied with an equal growth in the amount of information that average users get exposed to on a daily basis [9]. Literally, thousands of news stories are published online on a daily basis [10]. Daily news stories are not purely text-based; rather, they are often accompanied with multimedia elements such as images, videos and audio clips. Furthermore, blog posts and Social Media (SM) feeds are also considered to be news-oriented contents. These emerging news contents are becoming more associated with other more traditional news contents (traditional text-based news stories and articles) [11]. Hence, the variety and high accessibility levels of online news contents have transformed the Web to a huge news desk where users are literally spoiled for choice. A study by Bohn and Short [12] suggests that the online data levels are equivalent of each US citizen consuming 12 hours of information or media daily. As a consequence, users are overwhelmed with an indigestible amount of information on a second-by-second basis.

Furthermore, the emergence of news syndication technologies such as Rich Site Summary (RSS) and Extensible Markup Language (XML) allowed for more effective means for news aggregation and distribution to end-users [13]. However, online users are typically interested in reading those news stories that they are interested in [14]. Nonetheless, the amount of news contents produced everyday defies the abilities of users to keep abreast of their areas of interest. Such a user need has given rise to a variety of news analysis and monitoring systems. Such systems aimed at facilitating the news monitoring, aggregation and analysis tasks that average users need to perform.

The sheer amount of online news contents and the need for effective user tools for scanning and filtering them led to the advent of software solutions that aimed to serve that very purpose [15]. Traditional news analysis solutions primarily aim at helping online users gist through the news contents that they are interested in. However, it is noticeable that such systems mainly focus on straightforward crawling and analysis of news contents (e.g Meltwater Group and AP news aggregation services and tools [16]). Such news analysis systems depend on a model that is based on an incremental database of news contents. What is common in these systems is that they lake the analytical depth that users may look for manifested by the dynamic nature of online news.

Furthermore, traditional news aggregation and analysis systems usually lake the capability to predict the possible development paths of news stories. For instance, how a news story will develop in terms of its coverage patterns based on the historical instances in which similar stories were covered.

Moreover, more often than not, users often find themselves in a situation where they need to choose the news sources that provide the best coverage of their preferred news topics (for example, the best news source that covers technology news). This task is not straightforward without accurate tools that can provide a trusted analytical view on the best news sources according to topic/issue. These aspects of news analysis are becoming more significant especially with the Web's ability to be a 24/7 news service for online users.

This paper proposes the Visual News Screener (VNS) which is a news source analysis, prediction and visualization system. VNS will provide online users with effective tools for following the news stories that they are interested in while giving them different analytical insights about them. Furthermore, VNS will combine textual and visual analysis within a consolidated platform that can act as an advanced analytical predictive news screener. VNS will also be able to assess and evaluate the efficiency of news sources in covering certain news topics and issues.

The contribution of this paper is threefold, first it presents a novel system for news monitoring with a multi model approach to handle the very nature of live news feeds. Second, VNS will uniquely utilize Machine Learning (ML) and Business Intelligence (BI) algorithms to predict the possible development paths of any given news topic/issue. This feature will be based on the comprehensive analysis of a representative set of training data. Third, VNS will employ Quantitative Analysis as well as Text Analysis to assess the efficiency of online news sources in covering news stories according to topic or issue. This facility will aim to help the end-users choose the news sources that will provide the best coverage of their news topics/issues of interest. Ultimately, each news source will have a unique hallmark according to topic, which will be used for efficiency assessment as well as news source comparison purposes.

2. RELATED WORK

News analysis systems use different methods for aggregating, analyzing and archiving news contents. In relation to news contents analysis, there are a number of approaches adopted by a variety of news analysis platforms. These varied approaches are manifested by the underlying goals and objectives of the system in question [5]. Different systems focus on different areas of analysis according to the user needs. The main analysis elements in news analysis systems revolve on text, image, video and audio analysis. Prominent news analysis systems focus on one or more of the aforementioned analytical angels. However, in most of the notable systems, there is an emphasis on text analysis due to the relative ease of extracting patterns/trends and information from the analyzed textual data [17].

When talking about textual contents analysis in particular, it can be observed that there are many systems that focus on this type of news analysis. What is common between these systems is their attempt to extract patterns and relationships among the analyzed news contents. Sentiment analysis systems are a good example of the utilization of textual analysis algorithms to extract meaningful facts from the analyzed news [18]. Such systems employ the principles of Natural Language Processing (NLP), Computational Linguistics (CL) and text analytics to extract the positive, neutral or negative opinions about any given topic [19]. Examples of news sentiment analysis systems include the systems developed by [18], [20] and [21].

News text analysis can serve other purposes as well. For example, the system proposed by [22] focuses on pure textual data analysis of news articles to produce summaries of the analyzed news stories. Another noteworthy example is the system outlined by [17], which exploits Distributed Artificial Intelligence (DAI) techniques to complement quantitative financial information extracted from online sources. Another example is Lydia, which is a system for "Large-Scale News Analysis" [23] that emphasizes on pure text analysis to track "temporal and spatial distribution of the entities in the news" [23].

It should be noted within this context that "Text Summarization" [24] techniques play an important role in modern news analysis systems. Text summarization process works by reducing each news article to a selective set of the most essential facts [25]. These facts are usually established prior to the actual analysis process. The end result here is represented in the form of a few quantitative variables that can be analyzed more straightforwardly than analyzing pure text. For example, [26] use a "Vector Space Model" to produce summaries of news stories.

It is worth mentioning that there is a bigger tendency to incorporate video and image analysis within the contemporary news analysis systems. This increasing tendency can be attributed to the repaid advancements that video and image processing technologies have witnessed over the last few years [27]. Such developments allowed for the extraction of more meaningful and useful information form the analyzed multimedia-oriented contents. An example of a news analysis system that focuses on video analysis is the system outlined by [28]. Another example is the system developed by Salton [29], which focuses on text analysis in combination with small-scale image analysis.

Despite the rapid advancements that news monitoring and analysis systems have witnessed, they still suffer from a number of shortcomings. One of the drawbacks of such systems is their tendency to focus on the news aggregation side of functionality. Other aspects that may be equally important may be overlooked. Such aspects include the analysis of any given source’s efficiency in covering news stories according to the news story topic or issue. Another shortcoming is not providing visualized analytical view about the news issue/topic being viewed. For example, the coverage patterns of similar news stories in the past. VNS will aim surpass these shortcomings by furnishing its users with added-value services. These services will give the users a more insightful look at the news topics and issues that they are interested in. they will also enable them to make more informed decisions about the best news sources to follow as further detailed below.

VNS will focus on a number of areas in relation to news filtering and analysis. VNS focus areas will mainly revolve on text analysis, predictive analysis of news stories, analytical analysis and comparison of news sources and news visualization. These implementation aspects of VNS are further detailed in the subsequent sections of this paper.

3. IMPLEMENTATION METHODOLOGY

VNS will be devised for the analysis of news feeds whether they are traditional news feeds or SM feeds. VNS will provide an interactive dashboard (the News Screener (NS)) through which it will be possible to monitor how the news stories of interest are evolving/will evolve overtime. Furthermore, VNS will provide its users with the unique feature of assessing the efficiency of news sources in covering news stories by topic or issue. This feature will depend on a combination of quantitative and text analysis techniques to evaluate the ways by which news sources cover a specific news topic or issue.

The generic architecture of VNS will depend on a modular Service Oriented (SO) approach. The adopted SO model will be based on a number of specialized code modules that will provide the required functionality. VNS will operate on top of a lightweight database model that will not aim at building a comprehensive news archive. Rather, it will analyze the most relevant parts of news stories (heading, opening and closing paragraph) to get the required insights and visualizations (3D and 2D visualizations). The aggregated news contents will be formatted into standard XML files that will be used for retrieval and presentation (through XSLT or visualization APIs) purposes.

VNS will adopt a "Multi Model" approach to achieve its news analysis functionality. Each analytical feature will have an associated model in the form encapsulated system classes to handle it. In this model, news data is extracted from a group of predefined set of sources at three different levels: (i) news story title; (ii) news story subject (e.g Global Warming, iPhone SE, etc); (iii) news story opening and closing paragraphs. After aggregating the details of the concerned news story, it will undergo a filtering and classification (categorization) process. In this process, VNS will detect the keywords, category, topics and issues that are associated with the extracted news story.

VNS will adopt a different approach in relation to news story summarization compared to other similar systems. VNS will assume that the gist of any news story is available in the opening and closing paragraphs. This assumption best suits VNS's focus on fast and accurate analysis of live online news sources. Therefore the opening and closing paragraphs of any of the aggregated news stories will be analyzed to create the news story summary. This summary will be later used in the visualization, comparison and prediction processes.

3.1. Text Processing

One of the core elements of any news analysis system is the approach that it adopts to carry out the initial processing of the news contents that it aggregates. VNS will follow the approach of "Real-valued Vector Space Model" as highlighted by [29]. This approach is adopted because it is one of the most straightforward approaches for text analysis which employs spatial proximity for semantic proximity [17]. Text processing is preceded with initial text filtering. This filtering process involves removing textual noises [17] such as symbols and stop words. This is to ensure that the accuracy of the text processing and classification processes are as accurate and fast as possible.

Based on the adopted Real-valued Vector Space Model, each news story in VNS will have a set of features that correspond to its actual textual contents. Hence, each news story will be treated as a multidimensional entity. Each dimension will hold a set of attributes related to one of the story's features. Therefore, the words contained within the news story will be represented as vectors in a dimensional space. Consequently, the analyzed news story will eventually be the sum of the vectors that represent the terms used in the story [30]. Based on that model, each news story will have a words (terms) vector and the weight of each word within the document will be calculated by its frequency. For example, if we have a news story D it will have a "terms vector" as follows:

$$\vec{T} = \{Term1, Term2, Term3, \ldots. TermN\}$$

Consequently, each news story is conceptually represented by a matrix that represents the terms found in that specific story as follows:

$$D = T\ X\ N$$

The frequency of a certain term (for example T1) within the story will be denoted by F (T1). VNS will only consider the terms with frequencies that indicate their significance (i.e frequencies that are not too high (common articles such as: and, ore, the, etc) nor very low). In this instance, if we have a story D, it will have multiple feature vectors depending on the frequency of the terms (which represent actual news issues/topics) that it contains. Hence, each news story may have multiple feature vectors as follows:

$$\vec{A} = \{F1, F2, F3, \ldots..Fn\}$$

Based on the identified feature vectors, the news story will be categorized according to VNS's predefined set of news categories as further explained below.

3.2. News Categorization

One of the baseline considerations of VNS is the news classification it is going to use to segment the news contents that it is going to monitor and analyze. VNS will adopt a news classification system that will be based on a selected set of news categories (politics, economy, sports, etc). However, for the purpose of assessing the coverage efficiency of news sources, another dimension will be added to this classification system which is topic. The rationale behind adding topic is the requirement to assess the efficiency in which any given news outlet covers a certain type of news (e.g natural disasters). For the purpose of the proof-of-concept version of VNS, the news topics will be broadly divided into six main categories. Under these categories, a number of sub-categories that represent a selective set of news stories will exist. Table 1 illustrates the news categorization system that will be adopted by VNS. It is assumed that each category will have multiple sub levels. For illustration purposes, one sub level is shown followed by the list of news issues that may be associated with any given news topic.

Table 1. News Topic Classification

Category	Topics	Issues
Politics	▪ Civil War	▪ Syrian War ▪ Libyan war ▪ Iraq War ▪ etc
	▪ Elections	▪ American Presidentials ▪ EU Elections ▪ etc
	▪ Governments	▪ New Saudi Government ▪ New French Prime Minister

		▪ etc
	▪ Political Crisies	▪ Yemen Unrest ▪ Egypt Riots ▪ etc
	▪ Terrorism	▪ 9/11 ▪ Paris Attacks ▪ etc
	▪ The United Nations	▪ The Security Council ▪ Article 7 ▪ etc
	▪ etc	▪ etc
Economy	▪ Oil Prices	▪ OPEC Meeting ▪ Price wars ▪ etc
	▪ Recession	▪ Global Recession ▪ etc
	▪ Investment	▪ Projects ▪ FDI ▪ etc
	▪ The World Bank	▪ New projects ▪ Members ▪ etc
	▪ International Monetary Fund	▪ Memberships ▪ Loans ▪ etc
	▪ etc	▪ etc
Sports	▪ Football	▪ The Premiere League ▪ Lionel Messi ▪ etc
	▪ Basketball	▪ NBA ▪ The World Cup
	▪ etc	▪ etc
Technolgy	Smart Phones Security etc	iPhone Android etc

The actual process of filtering the news stories – which takes place after aggregating the monitored news stories from their designated sources - is threefold. Firstly, the scanned news stories are categorized according to the predefined news categories as shown in Table 1. In this process, the keywords in the title, opening and closing paragraphs of the scanned news story are compared with a predefined words repository. This repository will contain the most common keywords that are associated with the news categories that VNS monitors. Then, the frequency and density of the keywords are evaluated. Based on that, a news story is associated with one or more categories. In this context, we assume that the news categories are represented in a class set denoted with C:

C = {Politics, Economy, Sport, Weather, Crime,…..}

Each category is a class on its own right that consists of multiple subclasses (news topics). Each subclass represents a certain news topic. For example:

Politics = {Civil War, Elections, Governments,…….}

Based on the adopted news category anatomy, each topic will then have a set of news issues that are associated with it. If we denote any of the aggregated news stories with D, this news story is assigned to the closest news issue that it represents (based on similarity analysis), for example:

$$D \in \{Iraq\ War\}$$

In many instances, a news story D may belong to more than one news issue, for example:

$$D \in \{Iraq\ War\}\ AND\ D \in \{Syrian\ War\}$$

It is worth noting here that the similarity between the news story and its closest issue/s will be measured by a cosine measure as applied by [23]. For instance, to measure the similarity between a certain news issue X and a given news story D, the following formula will be followed:

$$sim\ (T, D) = \frac{\vec{Y}.\vec{Z}}{\|\vec{Y}\|.\|\vec{Z}\|}$$

where Y and Z represent two feature sets that VNS compares to create clusters of similar news stories. These clusters can be later used in the analysis, prediction and visualization processes as further elaborated below.

3.3.News Prediction

In addition to the basic news aggregation, filtration, classification, analysis and visualization functionality that VNS will support, it will also provide its users with the ability to view predictions about the expected development path/s of any news issue. This functionality will enable the users to have a unique perspective on their topics of interest.

The prediction approach that VNS will adopt will be based on the comprehensive analysis that it will conduct on a selected set of training data. This training data will represent the history of a number of selected news issues and topics from different sources. Different criteria about each issue will be analyzed. These criteria will include: (i) the number of news sources that covered the issue from the day it emerged (ii) the number of news stories published about that specific news issue (iii) the development patterns of the number of news stories published about that specific issue over time (increase, decrease or no change) (iv) the variety of sources that are covering the concerned news issue (for example, when will it pick momentum in social media).

As the case with any predication-based software application, the accuracy of the prediction results will improve over time with the effective employment of Business Intelligence (BI) and Machine Learning (ML) techniques. The prediction results will be presented to the end user graphically through a set of charts and graphs. Moreover, users will also be able to compare two or more news stories in terms of their possible development paths.

An example about VNS's expected prediction capability is the ability of predicting the development patterns of a news story within Twitter. For example, if we consider a news story about "People's Choice Award", the historical development of this story can be seen from Figure 1 based on the number of tweets related to it per month. VNS is ideally expected to provide users with a similar output about the future of this news issue based on its analysis of its historical data.

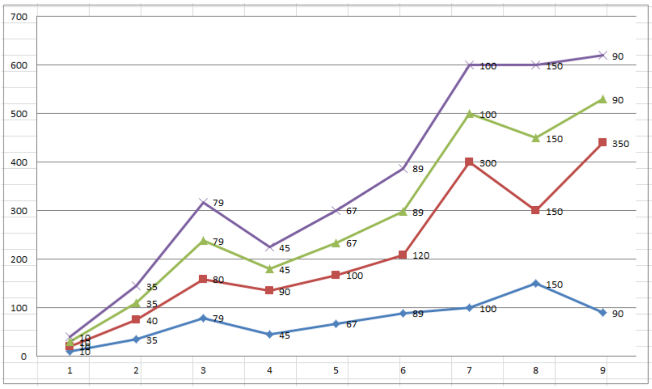


Fig 1: Development Paths of a News Issue in Twitter

3.4. Source Efficiency Analysis

Another unique functionality dimension of VNS is its ability to assess the efficiency of news sources in covering specific news topics and issues. This feature will aim to help the end-users follow the news sources that will provide the best coverage for their areas of interest. Similar to news prediction, this feature will depend on the analysis of a comprehensive set of training data. The training data will determine the efficiency of online news sources in covering certain news Topics (e.g Sports) or issues (e.g. American Presidentials). The actual analysis process will be based on a combination of quantitative and text analysis of the aggregated news stories.

News coverage efficiency analysis will depend on certain characteristics. Firstly, it will rely on the historic count of the news stories published about a certain topic/issue by a given online news source. Secondly, VNS will scan for certain keywords within the news it aggregates to determine whether the news source published a story of its own or a story quoted from another source. Thirdly, the number of times a certain source is quoted by other sources will also be used as a factor that determines the strength of this news source. Fourthly, VNS will look for other factors that indicate news coverage efficiency including the inclusion of multimedia elements (images, audio and video) as well as expert quotes and interviews. In addition to the factors above, an extra dimension will be added in the form if user voting. Hence, users will also be able to have their say on the news sources that they think provide the best coverage for their topics/issues of interest.

Based on the above criteria, VNS will be able to build a database about the coverage patterns of the online news sources that it will cover. Overtime, it is expected that VNS's ability to assess the efficiency of news sources will improve as the set of its training data grows. Furthermore, VNS will be able to have what we call "News Source Coverage Hallmark". This hallmark will be a numeric measure that is associated with each news source according to the topics or issues that it covers. This measure can be used for multiple purposes in the subsequent versions of VNS. For example, it can be used for comparing the coverage patterns of news sources. It is also planned to use it as a tool to determine the source of news stories based on the identified coverage patterns. Such features will make VNS a practical application that can guide its users to the best news sources to follow to keep abreast of their areas of interest.

4. SYSTEM SPECIFICATIONS

The general architectural approach of VNS is illustrated in Figure 1. The adopted architecture is based on an SO loosely-coupled model. VNS's database will be constantly fed with news feeds (XML, RSS, SM, Customized Feeds, Database Feeds, etc) that are aggregated by customized crawlers and news parsers. Following the adopted SO approach, message passing will be the basis on which the modular VNS components will communicate with each other.

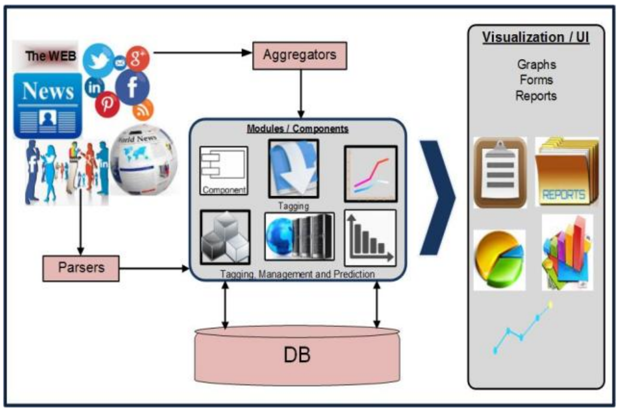
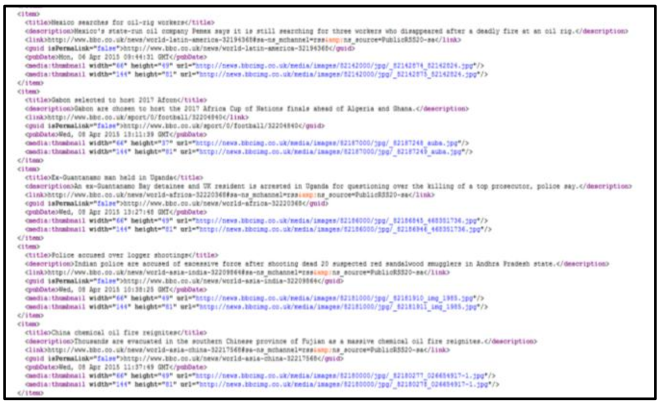


Fig 2: Generic VNS Architecture

An example of the news feeds that VNS will support are RSS feeds which are very common in news websites. Figure 3 illustrates a typical RSS feed that can be consumed by VNS.



```
<item>
  <title>Mexico searches for oil-rig workers</title>
  <description>Mexico's state-run oil company Pemex says it is still searching for three workers who disappeared after a deadly fire at an oil rig.</description>
  <link>http://www.bbc.co.uk/news/world-latin-america-32194368#sa-ns_mchannel=rss&amp;ns_source=PublicRSS20-sa</link>
  <guid isPermaLink="false">http://www.bbc.co.uk/news/world-latin-america-32194368</guid>
  <pubDate>Mon, 06 Apr 2015 09:44:31 GMT</pubDate>
  <media:thumbnail width="66" height="49" url="http://news.bbcimg.co.uk/media/images/82142000/jpg/_82142874_82142824.jpg"/>
  <media:thumbnail width="144" height="81" url="http://news.bbcimg.co.uk/media/images/82142000/jpg/_82142875_82142824.jpg"/>
</item>
<item>
  <title>Gabon selected to host 2017 Afcon</title>
  <description>Gabon are chosen to host the 2017 Africa Cup of Nations finals ahead of Algeria and Ghana.</description>
  <link>http://www.bbc.co.uk/sport/0/football/32204840</link>
  <guid isPermaLink="false">http://www.bbc.co.uk/sport/0/football/32204840</guid>
  <pubDate>Wed, 08 Apr 2015 13:11:39 GMT</pubDate>
  <media:thumbnail width="66" height="37" url="http://news.bbcimg.co.uk/media/images/82187000/jpg/_82187248_auba.jpg"/>
  <media:thumbnail width="144" height="81" url="http://news.bbcimg.co.uk/media/images/82187000/jpg/_82187249_auba.jpg"/>
</item>
<item>
  <title>Ex-Guantanamo man held in Uganda</title>
  <description>An ex-Guantanamo Bay detainee and UK resident is arrested in Uganda for questioning over the killing of a top prosecutor, police say.</description>
  <link>http://www.bbc.co.uk/news/world-africa-32220368#sa-ns_mchannel=rss&amp;ns_source=PublicRSS20-sa</link>
  <guid isPermaLink="false">http://www.bbc.co.uk/news/world-africa-32220368</guid>
  <pubDate>Wed, 08 Apr 2015 13:27:48 GMT</pubDate>
  <media:thumbnail width="66" height="49" url="http://news.bbcimg.co.uk/media/images/82186000/jpg/_82186845_468351736.jpg"/>
  <media:thumbnail width="144" height="81" url="http://news.bbcimg.co.uk/media/images/82186000/jpg/_82186846_468351736.jpg"/>
</item>
<item>
  <title>Police accused over logger shootings</title>
  <description>Indian police are accused of excessive force after shooting dead 20 suspected red sandalwood smugglers in Andhra Pradesh state.</description>
  <link>http://www.bbc.co.uk/news/world-asia-india-32209864#sa-ns_mchannel=rss&amp;ns_source=PublicRSS20-sa</link>
  <guid isPermaLink="false">http://www.bbc.co.uk/news/world-asia-india-32209864</guid>
  <pubDate>Wed, 08 Apr 2015 10:38:25 GMT</pubDate>
  <media:thumbnail width="66" height="49" url="http://news.bbcimg.co.uk/media/images/82181000/jpg/_82181910_img_1985.jpg"/>
  <media:thumbnail width="144" height="81" url="http://news.bbcimg.co.uk/media/images/82181000/jpg/_82181911_img_1985.jpg"/>
</item>
<item>
  <title>China chemical oil fire reignites</title>
  <description>Thousands are evacuated in the southern Chinese province of Fujian as a massive chemical oil fire reignites.</description>
  <link>http://www.bbc.co.uk/news/world-asia-china-32217568#sa-ns_mchannel=rss&amp;ns_source=PublicRSS20-sa</link>
  <guid isPermaLink="false">http://www.bbc.co.uk/news/world-asia-china-32217568</guid>
  <pubDate>Wed, 08 Apr 2015 11:37:49 GMT</pubDate>
  <media:thumbnail width="66" height="49" url="http://news.bbcimg.co.uk/media/images/82180000/jpg/_82180277_024654917-1.jpg"/>
  <media:thumbnail width="144" height="81" url="http://news.bbcimg.co.uk/media/images/82180000/jpg/_82180278_024654917-1.jpg"/>
</item>
```

Fig 3: Example of RSS Feed

4.1. Functional Modules

Based on the adopted SO approach, VNS will consist of eight core specialized functional modules as follows:

A. The News Crawler

This component will be responsible for crawling a number predefined news sources to extract the news contents that are of interest to the user. The crawler will be one of the channels through which contents are fed into the system. Regular Expressions (RegEx) in addition to customized parsing modules will form the functional basis of this module.

B. The News Aggregator

The different news feeds that the system will monitor and analyze will be aggregated by this module. The functionality of this module will complement the crawler module by allowing for the inclusion of news contents from external feeds (RSS and XML) as well as Application Programming Interfaces (APIs).

C. The News Predictor

This analytical component will predict the possible development paths of certain news topics and issues (the potential to further develop, freeze or decline). A number of customized analytical algorithms (Regression, Association, etc) will be adopted to produce the predictions based on Business Intelligence (BI) and training data that will be accumulated over time.

D. The Topic Manager

The news topics/issues that the system is monitoring will be managed through this module. News will be collected, tagged and grouped according to the actual topics that users are interested in.

E. The News Screener

This module represents the User Interface (UI) part of VNS. It consists of a highly customizable web-based frontend that displays the different news feeds and indicators. Users will be able to choose from a set of visualization and monitoring options as further detailed in the operational scenario below.

F. The News Manager

The administrative backend of VNS is represented by this component. Through this administrative backend it will be possible to specify the news topics, the feeds to include and the visualization options to be presented to end-users. The different elements of the News Screener will also be managed through this module.

G. The News Archiver

This module will aim to provide the historical context of the news topics of interest. The News Archiver will be a backbone component through which it will be possible to have more accurate analytical results based on the data that it will accumulate over time. Moreover, the News Archiver will contain most of the training data that will be used by the News Predictor and Source Analyzer Modules.

I. The News Visualizer

This module will provide the different visualization options that can be applied to the analyzed news. Contents will be passed to the News Visualizer in the form of well-structured XML files that are ready to be visualized within the News Screener module. Different visualization techniques will be supported such as Line Graphs, Time Series Graphs, Histogram, Scatterplot, Area Charts, etc.

J. The Source Analyser

This module will be responsible for assessing the efficiency of news sources in covering certain news topics/issues. The operation of this module will depend on an analytical model. This model will assess certain criteria in relation to the coverage patterns of the covered news sources (see Section 3.4 for more details about the assessment criteria).

The abovementioned modules are illustrated in Figure 4. It can be seen that VNS is adopting a pure SO modular approach where the concerned modules communicate with each other to achieve the required functionality. Internal module communication is based on message passing as functional requests and data are constantly flowing between VNS's modules. The actual operational model of VNS is further highlighted in the illustrative operational scenario in Section 5.

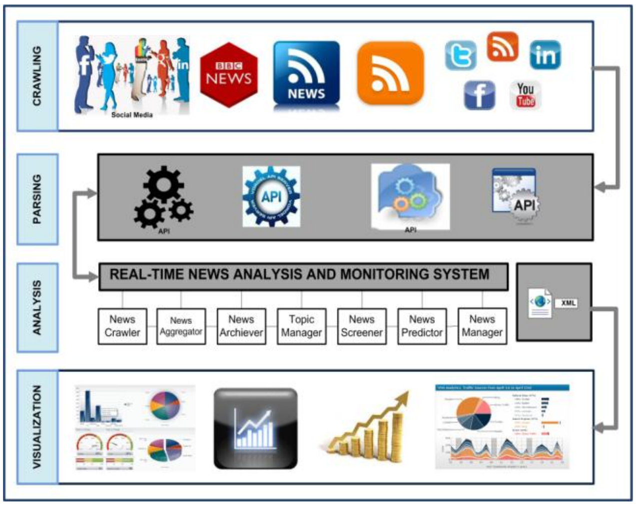


Fig 4: VNS Modules

5. OPERATIONAL SCENARIO

Online users are usually interested in different aspects of news monitoring and analysis. A typical case study that depicts the benefits of VNS is represented in the scenario whereby a user has a long-term interest in a certain news topic or issue. In such a case, the user will be inclined to know three main aspects of the topic/issue of interest. First, he will be interested to know the current status of news coverage in relation to that topic/issue. Second, it will be intriguing for that user to know how that news issue of interest is going to evolve over time in terms of the depth and breadth of its coverage. Third, it will be useful for the user to follow the news sources that will provide the best coverage for the topic/issue that he is following. VNS will empower users to perform the above tasks as further detailed in the operational scenario below. This scenario will be outlined in conjunction with detailing the different functional components of VNS.

A typical operational scenario within VNS will involve the process of monitoring one or more issues of interest from a number of online news sources. In the illustrative operational scenario highlighted below, it is assumed that the user is interested in the news related to the release iPhone SE. The steps followed in this scenario will be as follows:

A. Topic Definition and Source Selection

In this step the user will use the "News Manager" module to define the keywords that are associated with the topic he is interested in. In our scenario the entered keywords will be as follows: "iPhone SE, Apple iPhone SE, iPhone SE Release". After specifying the keywords, the user will proceed to choosing the preferred sources from which the related news contents will be fetched. VNS interface will have a predefined list of news sources and SM feeds. The user will be able to choose any combination of sources.

B. News Crawling and Parsing

A customized web crawler will be used to parse the news contents related to the news issue of interest. VNS crawler will primarily use the technique of "Regular Expressions" (RegEx) which is an effective programmatic approach for parsing a selective set of web contents [31]. RegEx will be complemented with a special algorithm for reading the actual contents of the opening and closing paragraphs of the crawled news stories (depending mainly on Bayesian filtering). The programmatic backbone of the crawler will be based on

WebSPHINX, which is an open source RegEx-base crawler [31]. For VNS, WebSPHINX will be customized and improved to meet the contents extraction needs of the system. The parsed contests themselves will be formatted as valid XML files that will be used in the archiving, analysis and visualization processes. The workflow of the crawling process is illustrated in Figure 5.

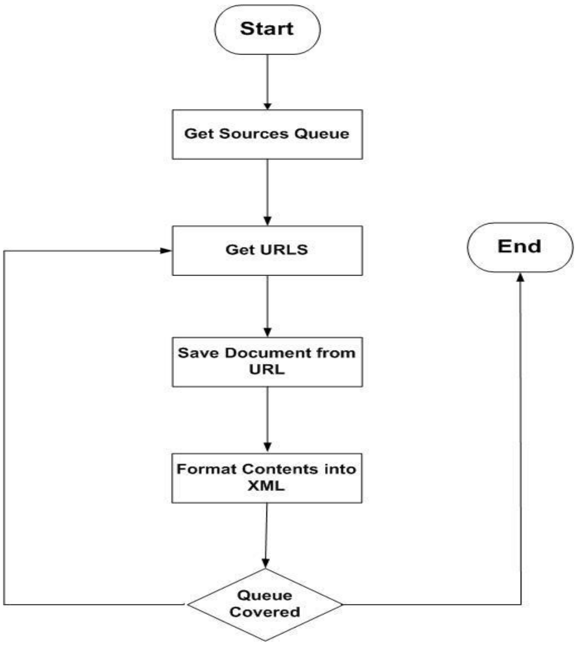


Fig 5: News Crawling Workflow

It can be seen from Figure 3 that a queue of preselected URLs will be crawled to index their news contents. Different Regular Expressions will be applied depending on the structure of the web page being crawled as shown in this example code excerpt:

```
$regexp = "<a\s[^>]*iPhone6(\"??)([^\" >]*?)\\1[^>]*>(.*)";
```

Once contents are parsed they will be filtered to eliminate any "noise" in the parsed contents. The parsed contents will then be formatted into well-structured XML files that will then be used for the monitoring, archiving, retrieval and visualization processes. In our iPhone SE example, a typical XML file structure will be formatted as follows:

```
<news_item_10>
<title>Apple iPhone SE pre-orders hit record 4 million on first day</title>
<date>Mon, 15 Sep 2015 16:34:40 GMT</date>
<source>Reuters</source>
<abstract>Apple Inc said many customers will need to wait until next month for their new iPhones after a record 4 million first-day pre-orders were logged, double the number for the iPhone 6 two years ago.</abstract>

<source_link>http://www.reuters.com/article/2014/09/15/us-apple-iphone-idUSKBN0HA1A220140915?feedType=RSS&feedName=technologyNews</source_link>

</news_item_10>
```

C. News Contents Tagging

News contents will be tagged based on the keywords that they contain. Each news story will be automatically tagged by employing a smart "Text Mining and Labelling" algorithm. This algorithm will identify the keywords and phrases that exist in each indexed item.

D. News Analysis

The news analysis will aim to identify the following criteria:

- Number of news stories/contents related to the topic/issue of interest now and in the past
- The trends pertaining the topic of interest within the identified sources
- Any abnormalities or unusual trends
- Predicting the possible development paths of the news issue

For each of the purposes above, a variety of analysis techniques will be employed as shown in Table 2 below.

Table 2. The Algorithms that will be used in VNS

Purpose	Algorithm Type
News Count	• Regression algorithms
Main Trends	• Association algorithms • Sequence analysis • algorithms
Abnormalities	• Data mining algorithms • Heuristics and calculations
Prediction	• Sequence Clustering Algorithms • Association Algorithms • Clustering Algorithms

E. News Visualization

The results of the news analysis process will be shown through a website frontend which will constitute the UI elements of VNS. The visualization process will depend on the actual outcome of analysis process. For example, predictions will be visualized as histograms and line graphs.

F. Source Selection

Through VNS's ability to assess the efficiency of news sources in terms of their coverage patterns of a certain news topic/issue, users will be able to choose the sources that provide the best converge for the release of iPhone SE. This selection will be based on the sources that VNS will recommend based on its assessment of the news sources defined within the system. This assessment process will be based on the assessment criteria explained in section 3.4. After choosing these sources, users will be able to perform the news monitoring, visualization and prediction tasks highlighted above.

6. SYSTEM IMPLEMENTATION

In order to showcase the feasibility and applicability of the proposed system, it was implemented programmatically in the form of a dynamic web application. The .NET Framework [32] was employed as the development platform for the proof-of-concept system. This system was built based on VNS's proposed functionality and features. However, VNS can be implemented in any suitable web technology (for instance: Java or PhP). VNS can also be built as a conventional desktop application depending on the actual user needs and requirements.

The core architectural approach that VNS is adopting is based on an SO approach. In this approach, a number of specialized code components collaborate to achieve the desired system's functionality. The main design pattern that was chosen to implement VNS is based on the Model-View-Controller (MVC) [33] design pattern. This design pattern was adopted as it allows for creating applications that totally separate the Database, Business Logic and User Interface (UI) elements [34]. Hence, this design pattern suited VNS's SO nature. Adopting MVC also allowed for a flexible implementation framework that is adaptable and extensible.

6.1. Creation of VNS Components

Following the adopted SO approach, each of VNS's components (the News Crawler, the News Aggregator, the News Predictor, the Topic Manager, the News Screener, the News Manager, the News Archiver, the News Visualizer and the Source Analyser) was built as an encapsulated code class. These classes communicated with each other via message passing. Message passing was facilitated by the code infrastructure that the .NET Framework has provided [33].

Furthermore, VNS components interact with a database backend that represents the Model part of the adopted MVC design pattern. This database backend was created using MS SQL Server and provided the archival as well data management capabilities that VNS needed. As can be seen in Figure 6, the whole VNS system is relying on a database backbone that interacts with VNS's functional components (VNS's SO modules). Message passing between VNS's components based on the adopted loosely-coupled approach facilitated the performance of the different system functions (news crawling, analysis, visualization, etc).

The UI elements of VNS (the View) are totally separated and from the business logic and database elements (the Controller and Model elements). Moreover, VNS's UI can take several forms such as a classic website frontend, a mobile application or even a conventional desktop application if the user needs necessitate such an interface.

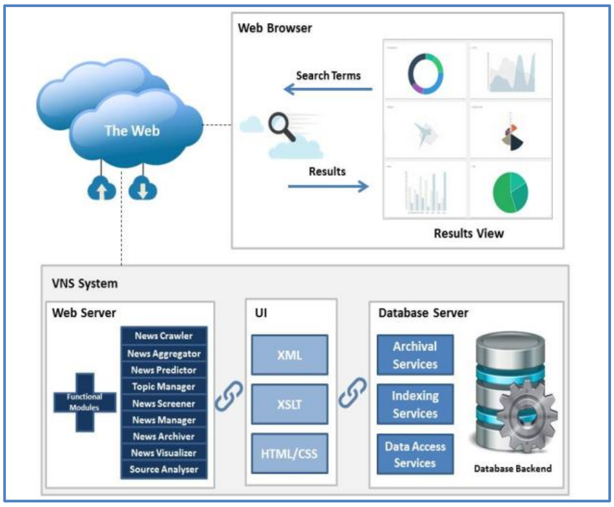


Fig 6: VNS Implmentation Components

6.2. Testing and Validation

To assess VNS's ability to perform its desired functionality, it was tested with the operational scenario that was highlighted in Section 5. Hence the following aspects of VNS were tested:

6.2.1. The User Interface

As mentioned above, VNS was created as a dynamic website application. VNS's UI is based on fully adaptive HTML5/CSS3 [35] website interface. This design approach was chosen so that VNS can be easily accessed via either PCs/Laptops or mobile devises.

Figure 7 illustrates VNS's dashboard where the user can view the latest trends and figures in relation to his preferred search terms. Users can also choose - from navigation links - to either: carry out a new search, manage the news sources or view the news visualizations.

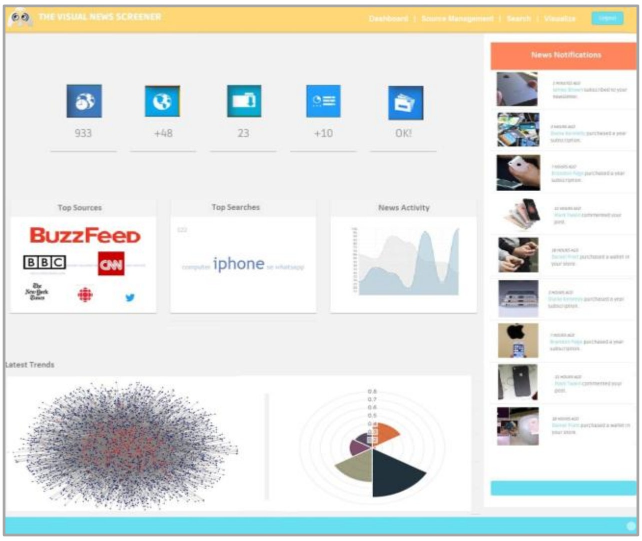


Fig 7: VNS Dashboard

6.2.2. News Contents Crawling and Indexing

As per the operational scenario adopted in this paper, a number of online news sources were used to parse the news stories which are related to "iPhone SE" and "iPhone" in general. This news issue formed the "test bed" on which VNS's functional points were tested and validated. In total, 100 online news sources were used for testing purposes.

These news sources included various online news outlets as well as Twitter feeds (as a representative agent for SM news feeds). Consequently, the news stories set included a total of 14,800 news stories which covered the last 5 years. These aggregated news stories also constituted the training data set. VNS used this training data for news source evaluation and well as news prediction purposes.

It was taken into consideration to cover the most prominent online sources so that the system can have amble training data to validate its functionality. Table 3 below lists some of the main online news sources that were used in the testing process.

Table 3. Some of the used news sources

Sample Source	Number of News Stories
CNN	2,850
NBC News Digital	1,150
BBC	21,000
The Huffingtonpost	9,130
CBS News	10,200
Buzzfeed	3,280
The New York Times	31,500
Mail Online / Daily Mail	20,300
Businessinsider	12,300
Twitter	9,000
Other Sources	20,000
Total News Stories	140,710

Based on the chosen news sources, news contents were crawled using VNS's News Crawler and Aggregator modules. These modules crawled the identified news sources and extracted the most relevant textual information (based on the search keywords). The metadata that is associated with each story (keywords, publication date, etc) was also extracted alongside each news story.

The news aggregation functionality itself was managed via a special administrative UI. This UI enabled users to define the target keywords as well as the news sources and feeds to be covered as illustrated in Figure 8. It can be seen that users are provided with the option of specifying their search keywords as well as the preferred sources. The crawling process progress is also shown in the form of a progress bar.

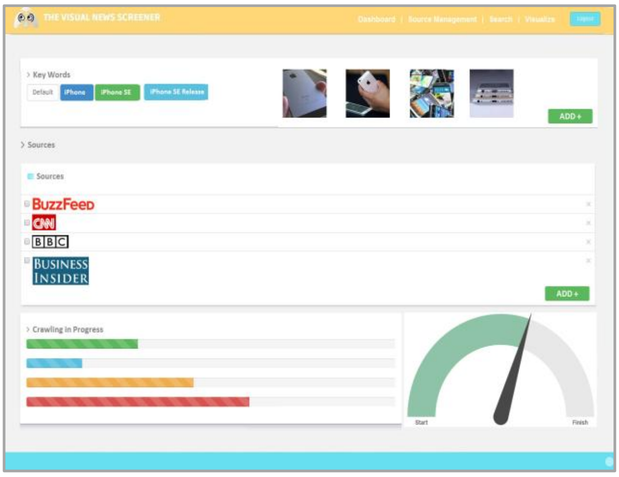


Fig 8: The Crawler Screen

6.2.3. Source Evaluation

The efficiency of news sources in covering certain news topics and issues was one of the most important distinctive features of VNS. This feature was tested based on the data that was collected from the news sources set. As explained in Section 4, certain criteria is used in assessing the efficiency of any given news source in covering a specific news topic or issue.

In our testing operational scenario, the efficiency of each of the covered news sources was assessed based on the accumulated data collected from each source. Consequently, each of the sources that were used in the system was evaluated based in the following criteria:

A. Number of News Stories

This measure is based on any source's ability to comprehensively cover any news topic or issue. This is a numerical measure that counts the total news stories that any given source has published about a given news issue. This formed the most basic efficiency measure.

B. Percentage of Original Stories

This is a numerical measure that represents the percentage of original stories (stories created by the news source) to the percentage of stories that the source is quoting from other sources. In this context, news story originality is another measure that is used to determine any given source's coverage efficiency. RegEx expressions were used here to determine whether a news story is original or not.

C. Number of Times a Source Quoted in other Sources

This is a numerical measure that counts the number of times a certain source is quoted/mentioned in other sources in relation to the news issue of interest. The more a source is quoted in other sources, the more efficient it is considered to be in its coverage patterns.

D. Precsne of Multimeida Elements

This is a numerical measure that counts the multimedia elements (images or videos) that are included in any of the aggregated news stories. Using multimedia elements is considered to be a user experience enhancer. Therefore, it is considered to be an efficiency measure. Specially-written RegEx statements were used to identify the multimedia elements that are embedded within each of the crawled news stories.

The criteria above were used to assess news source efficiency at a top level within the proof-of-concept VNS system. More detailed measures can be used in the future versions of the system. Such measures may include SM coverage, continuity of the news coverage; among other measures that can give a more accurate rating of each news source. Table 4 summarizes the analysis results of the main sources used in the test VNS system.

Table 4. Source Analysis Results

Source	Story Count	Original?	Quoted?	Multimedia
CNN Network	2,850	2,850	300	8,550
BBC	21,000	21,000	5,003	31,500
CBS News	10,200	7,666	6,733	17,327
Buzzfeed	3,280	2,009	1,271	2,007
The New York Times	31,500	31,000	17,500	55,000
Mail Online / Daily Mail	20,300	17,000	919	4,4011
BusinessInsider	12,300	11,003	801	13,444

Figure 9. Illustrates the interface through which VNS users can view the assessment results of the news sources. Relevant visualizations are shown in accordance to the chosen news topic or issue. In this example, the assessment results are visualized in the form of an Area Chart. This chart enables the user to identify the highest score source at a glance. In the context of the test search case (iPhone SE), the graph shows that the heights score source is The New York Times as the graph reaches its magnitude adjacent to that source.

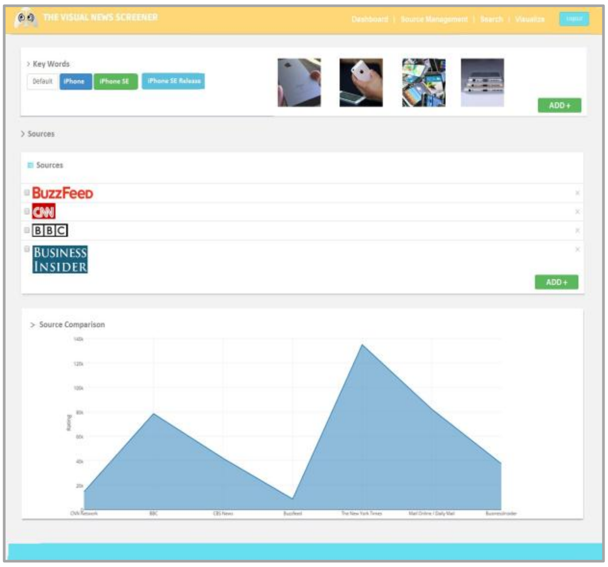


Fig 9: The Crawler Screen

6.2.4. Results Visualization

News visualization takes multiple forms within VNS's UI. Firstly, the coverage patterns and trends of news stories are visualized through graphical representations. Furthermore, news source efficiency comparisons are summarized in a set of charts that highlight the relative merits of each source as explained above.

Visualization within VNS can be used for multiple purposes. Firstly, it can be used for comparing different news topics or issues according to their coverage patterns and trends. Visualization can also be used to assess the efficiency of each of the covered news sources as explained before. Furthermore, graphs and charts can be a useful tool to get an insight about the historical coverage patterns of a certain news topic or issue.

As an example, different types of visualizations can be used to summarize the efficiency of each news source. Figure 10 shows how VNS is utilizing Irregular, Radial, Funnel and Stream graphs to visualize the efficiency rating of each news source.

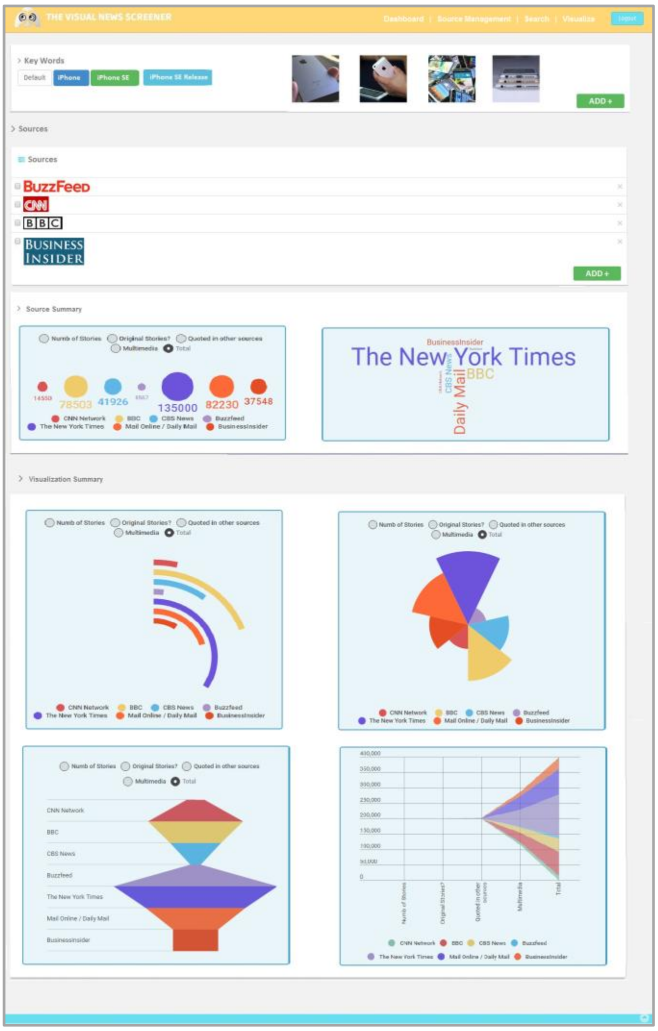


Fig 10: Different Visulization Options

VNS also allows the users to view the correlation between separate news topics or issues through Scatter charts. The chart example in Figure 11 highlights the correlation patterns between "Apple Sales" and "iPhone SE" news stories. The correlation dimension used here is the number of news stories published about each news issue. However, other dimensions can also be used such as number of SM posts, number of sources that are covering the story, etc.

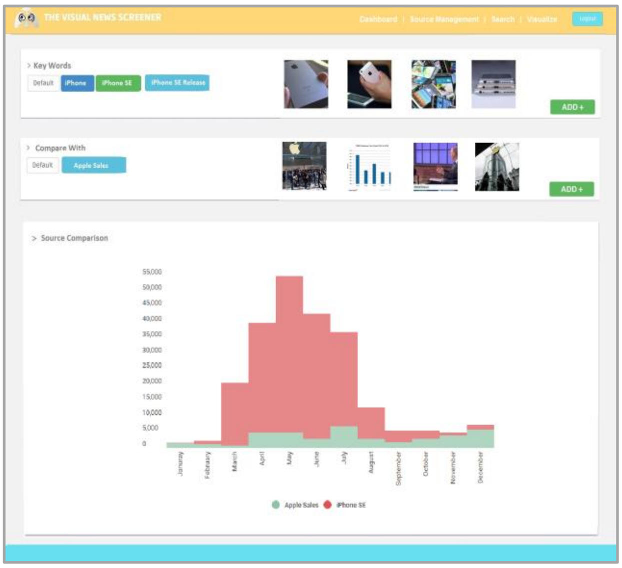


Fig 11: The Crawler Screen

7. CONCLUSIONS AND FUTURE WORK

This paper outlined a proposed system for news source analysis and visualization under the name VNS. VNS aims to provide its users with an insightful look at the current status and development paths of any news story. VNS goes beyond the functionality of traditional news aggregators by the source evaluation and visualization capabilities that it has.

VNS will also have a prediction function that will aim to predict the possible development paths of the monitored news topics. VNS will also be able to assess the coverage efficiency of news sources. Furthermore, the core VNS's UI will depend on news visualization where the analysis results will be graphically presented.

This paper highlighted the implementation of a proof-of-concept version of VNS. This version of VNS was built as an interactive web application. This VNS version was tested with an operational scenario that illustrated the data analysis and visualization capabilities of the system. The testing process was based on a set of comprehensive training data that was aggregate from live online news sources.

The future work will involve the actual implementation of the full-fledged VNS system. This implementation is planned to be based on the proposed architectural approach while experimenting with a rich set of training and testing data. VNS implementation will be followed by a testing period. In this testing period, different operational scenarios will be applied and evaluated. This step will be accompanied with testing the actual visualization, prediction and assessment algorithms that will form the basis of VNS's unique functionality. The final step will involve the system deployment via multiple mediums including web and mobile applications.

8. REFERENCES

[1] Norris, P. 2000. The Worldwide Digital Divide: Information Poverty, the Internet and Development, The Annual Meeting of the Political Studies Association of the UK London Harvard University.

[2] Mitchelstein, E. and Boczkowski, P. J. 2009. Between tradition and change A review of recent research on online news production journalism 562-586.

[3] Vasileios M., Spyros G., Georgios P., Walter K., Jorg S., Roeland O., Marijn H., and Franciska J. 2012. A System for the Semantic Multimodal Analysis of News Audio-Visual Content, EURASIP Journal on Advances in Signal Processing

[4] Allan S. 2006. Online News: Journalism and the Internet, Maidenhead Open University Press.

[5] Martin C., and John S. 2008. The Future of Newspapers, Journalism Studies 650-661.

[6] Flew, T. 2009. Democracy, participation and convergent media: case studies in contemporary online news journalism in Australia, Communication, Politics & Culture 87-109.

[7] Lebedev, E. 2016. The Independent, http://www.independent.co.uk/news/media/press/the-independent-becomes-the-first-national-newspaper-to-embrace-a-global-digital-only-future-a6869736.html.

[8] Sylvia O., Hyejoon R., and Amy Z. 2013. Mobile News Adoption among Young Adults Examining the Roles of Perceptions, News Consumption, and Media Usage, Journalism & Mass Communication Quarterly.

[9] Sriram K., and Shyam S. 2006. The Psychological Appeal of Personalized Content in Web Portals: Does Customization Affect Attitudes and Behavior?, Journal of Communication.

[10] Dragomir R., Jahna, O. and Adam W. 2005. NewsInEssence: summarizing online news topics, Communications of the ACM 95-98.

[11] Sitaram A., Bernardo H. 2010. Predicting the Future with Social Media, Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM 492 - 499 Toronto, IEEE.

[12] Hudson, A. 2012. BBC, http://news.bbc.co.uk/2/hi/programmes/click_online/9742180.stm.

[13] Gill, K. 2005. Blogging, RSS and the Information Landscape: A Look at Online News, WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics 120-122.

[14] Harumi M., Masao K., Hidekazu A., Atsushi S., Hideaki T., and Toyoaki N. 1997. Weak Information Structures for Community Information Sharing, international journal of knowledge-based and intelligent engineering systems 225-234.

[15] Chris P., and David D. 2009. Making Online News: The Ethnography of New Media Production, Journal of Information Technology & Politics 189-190.

[16] Quinn, J. 2014. Associated Press v. Meltwater: Are Courts Being Fair to News Aggregators? Minnesota Journal of Law, Science & Technology 1189-1219.

[17] Young-Woo S., Joseph G., and Katia S. 2004. Financial News Analysis for Intelligent Portfolio Management, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania.

[18] Namrata G., Manja S., and Steven S. 2009. Large-Scale Sentiment Analysis for News and Blogs, 3rd Int'l AAAI Conference on Weblogs and Social Media184-188San Jose, California Association for the Advancement of Artificial Intelligence.

[19] Nabeela A., Gaber M., and Mihaela C. 2013. SA-E: Sentiment Analysis for Education, 5th KES International Conference on Intelligent Decision Technologies 155-160, Sesimbra.

[20] Bo P., and Lillian L. 2002. Thumbs up?: sentiment classification using machine learning techniques, The ACL-02 conference on Empirical methods in natural language processing 79-86 Stroudsburg, PAACM.

[21] Nasukawa. J. Yi, T., Bunescu, R. and Niblack, W. 2003. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques, Data Mining, ICDM 2003. Third IEEE International Conference, 427-434 San Jose, CAIEEE.

[22] Takaharu T., and Atsuhiro T. 2008. News Aggregating System with Automatic Summarization, INFOS2008, Cairo-Egypt.

[23] Levon L., Dimitrios K., and Steven S. 2005. Lydia: A System for Large-Scale News Analysis, String Processing and Information Retrieval 161-166.

[24] Dipanjan D., and Andre F.T. 2014. A Survey on Automatic Text Summarization International Journal of Computer Science and Information Technologies 7889-7893.

[25] Luke M., Jinzhu J., Brian G., Bin Y., and Laurent G. 2011. Summarizing Large-scale, Multiple-document News Data: Sparse Methods and Human Validation Berkeley, CA University of California, Berkeley.

[26] Brian G., Jinzhu J., Luke M., Laurent G., Bin Y., and Sophie C. 2010. Discovering word associations in news media via feature selection and sparse classification, The international conference on Multimedia information retrieval 211-220 New York, NYACm.

[27] Lekha C., Tat-Seng C., and Chin-Hui L. 2003. A Multi-Modal Approach to Story Segmentation for News Video, World Wide Web 187-208.

[28] Qi, W. Gu, L., Jiang, H., and Chen, X. 2000. Integrating visual, audio and text analysis for news video International Conference on Image Processing 520 - 523 Vancouver, BCIEEE.

[29] Salton, G. 1989. Automatic text processing: the transformation, analysis, and retrieval of information by computer, Boston Addison-Wesley Longman Publishing Co., In BBC.

[30] Arvind N., Jeevan S., Alexandre P., and Andrew M. 2015. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector SpaceIthaca, NY Cornell University.

[31] Christian G., Filippo R., and Paolo T. (2006), Web crawlers compared, International Journal of Web Information Systems 85 - 94.

[32] Lapalme, J., Que., Aboulhamid, M., Nicolescu, G., and Charest, L. 2004. .NET framework - a solution for the next generation tools for system-level modeling and simulation, Design, Automation and Test in Europe Conference and Exhibition 732 - 733 Montreal, IEEE.

[33] Marston, T. 2004. The Model-View-Controller (MVC) Design Pattern for PHP, www./php-mysql/model-view-controller.html.

[34] LI Y. 2005. Improvement and Application of MVC Design Patterns, http://en.cnki.com.cn/Article_en/CJFDTOTAL-JSJC200509035.htm.

[35] Karl A., and Dan J. 2012. Mobile e-services using HTML5, Local Computer Networks Workshops (LCN Workshops)814 - 819Clearwater, Clearwater, FL, IEEE.