MAPNEWS: A Framework for Aggregating and Organizing Online News Articles

Jeelani Ahmed
Department of Computer Science and Information Technology, Maulana Azad National Urdu University, India
jeelani.jk@gmail.com

Muqeem Ahmed
Department of Computer Science and Information Technology, Maulana Azad National Urdu University, India
muqeem.ahmed@gmail.com

Abstract: *In recent years, digital news has become increasingly prevalent, with many people getting their news and information from online sources rather than traditional print or broadcast media. This shift has been driven, in part, by the convenience and accessibility of digital platforms, as well as the ability to personalize and customize news feeds. Digital news also allows for greater interactivity and engagement with readers and can reach a global audience almost instantly. News articles contain a plethora of hidden spatial information that, when shared with readers, increases comprehension of current events. Only a few news aggregation systems make this information available to users. Many stories, on the other hand, are not clearly geotagged with their spatial information. In this work, we propose the MapNews framework, a novel system that gathers, analyzes, and presents news articles on a map interface, allowing users to take advantage of their underlying spatial information. MapNews pulls content from several different internet news sources and, using a custom-built geotagger, it extracts geographic content from articles. A rapid online clustering method is used to organize articles into story clusters. Panning and zooming MapNews' map interface allows readers to receive news based on geographic location and category importance, and they will view distinct articles depending on their location. MapNews achieved an ARI score of 0.89 for clustering and an accuracy of 95% in usability testing.*

Keywords: *News aggregation, information retrieval, clustering, data aggregation, web scrapping.*

Received November 9, 2021; accepted December 14, 2022
https://doi.org/10.34028/iajit/20/3/10

1. Introduction

With the advent of the internet and social media, a diverse range of news sources has been available. Because these news sources are so easily accessible, a flood of information is generated, which is frequently paradoxical and misleading. Most news aggregators such as Microsoft Live News [29], Reddit News [35], Yahoo! News [41], and Google News [12], have just an elementary knowledge of the implied spatial context of news stories [3, 8], which is generally reliant on the publishing newspaper's location. Additionally, rather than grouping articles by location, these systems classify them by keyword or topic [22]. Given that readers' location-related characteristics drive most of their interest in news, it's surprising that they can't handle a couple of common prevalent forms of geography-related queries:

1. Feature-based: "What is the location of the news report?"
2. Geographic-based: "What is going on in a certain location?".

Our intention is to allow users to respond to the above questions by offering the reactions through a map interface instead of the traditional static interface similar to digital newsprint, which is usually linear and static, in which an editor determines the significance of news stories and display them accordingly in the order of relevance with no consideration of location. On the other hand, the map interface is dynamic, making it possible for the articles to be linked with a specific place and can change over time without affecting the placement of other news articles.

We introduce a real-time system named MapNews ("A web-based system for aggregating and displaying online news articles") to address the above questions. News articles are collected and based on their geographic and textual content MapNews organizes them into story clusters and intelligently links available spatial references with the news articles knows as geotagging. For example, a news item stating "Today, New Parliament Budget presented at New Delhi" is represented by a properly positioned location marker on the map at the position referring to New Delhi, India.

MapNews is built to be extensible, reactive, and modular, with article processing split amongst many modules (Segment 3). A transactional database system is at the center of the system, through which all modules communicate. As stated in (Segment 4), the system gathers and preprocesses news stories from a variety of online sources. The geotagger in MapNews (Segment 5) gives location coordinates for every news item, and then, using an online clustering method (Segment 6),

news items are subsequently clustered by category into story clusters. Articles are also aggregated by geography (Segment 7) and rated by article importance, determined by the quantity of different news sources that reference the news article.

Furthermore, depending upon the zoom level and current location in the map interface, news articles are geographically aggregated and sorted (Segment 8). For instance, readers can see markers corresponding to articles relevant to an international audience when watching the whole planet on the map, giving them an idea of what is happening around the world. MapNews continually refreshes the map when readers zoom in and pan across diverse geographical locations, ensuring that the display is always packed with appropriate article markers. Users may zoom in to see increasingly local news at the country, state, or city level. These effortless controlling features allow readers to efficiently comprehend up-to-date news events in terms of geography simply by extracting geographic information from news articles.

Our work is closely related to geographic information extraction, as the MapNews system uses news articles to retrieve geographic locations. Most researchers have focused on determining the geographic extent of individual documents and websites [2]. With regard to news stories, there may be three types of geographic scope:

1. Publisher Scope: the physical location of the publisher.
2. Reader Scope: the location of the readers.
3. Content Scope: the geographic location of the story content. In our proposed system, the geographic scope of an article is determined by its content.

Other techniques [10, 13, 20, 22, 27] take advantage of the article's inbound and outbound link structure.

The following are the main contribution of our work:

- Extraction of news articles from various news websites.
- Geotagged news articles to find their spatial information
- Clustering and aggregation of news articles with spatial data on a world map interface.

The rest of the article is organized as below: Segment 2 goes through the background information and similar work about news aggregation systems. Segment 3 proposed the MapNews system. Segment 4 discusses the algorithmic approach. Segments 5 and 6 explain the cluster focus and Map interface respectively. Finally, segments 7 and 8 concludes with discussion and a conclusion.

2. Related Work

Much of the existing work on news aggregation is based on a typical linear interface that looks like a newspaper, with items grouped and displayed in order of significance as determined by an editor, with no regard for geography [2]. Some of such news aggregation systems are:

2.1. Google News

Google News [19] is a news website that collects stories and reports from over 4,500 news sources. It can classify and show analogous material based on the user's preferences [12]. Google News has numerous sections on its main page, starting with 'Top Stories,' then 'Politics,' 'World News,' 'Sports,' and so on. These topics of news articles are listed in a vertical menu on the website's left side, with links to pages containing relevant material. Additionally, signed-in Google News users get access to extra features. Such as the ability to quickly access past media articles that the user has already browsed and, readers have the option of receiving news recommendations based on their search history [12]. Furthermore, Google News has a hybrid filtering mechanism, which means that suggestions are based on the user's search history and detailed ratings [32].

2.2. Microsoft News

Using the Microsoft Lives News app, you can read the most recent headlines and articles selected by the company's editors. The website covers a wide range of topics, including top headlines, global news, US, crime, sports, technology, and opinion. Users may customize their favorite categories and sources, receive breaking news alerts through alerts, prioritize desired news sources, and change font sizes to make articles simpler to read [1]. At the top of the website, all the news category menus are available, and then news stories are displayed in a traditional linear design with no geographic content [33].

2.3. Yahoo! News

Yahoo! started Yahoo! News which is an internet-based news aggregator news website. It aggregates news stories and reports from around the globe and displays them according to the user's preferences. Yahoo! News' home page has a variety of news topics, including Politics, World, Covid19, Health, Science, and Climate Change [24]. These categories are visible at the top of the website, linked to a page with material corresponding to each class. For signed-in users, Yahoo! News also provides relevant and suggested items. Given these considerations, Yahoo! News also lacks geographic material and interactive maps in its news stories.

MapNews is a unique and different system compared with the Yahoo! News, Microsoft Live News, and the popular Google News since MapNews uses a map interface as a medium for spatial news aggregation. In

contrast, most of today's news aggregators provide limited local news coverage that can generally be found by specifying an area pin code. The user's selection of news stories appears to be based mainly on the newspaper's publishing region rather than article data.

MapNews analyses terms that are probably related to geographic locations to retrieve an article's geographic focus [7]. In Natural Language Processing (NLP), there is a well-researched topic called Named-Entity Recognition (NER) [43] that recognizes things like people, places, and organizations. A range of methods are used by today's NER taggers, such as Natural language processing [11, 30, 34], statistical learning [6, 14, 25, 28] and hybrid approaches [31].

The geotagger at MapNews has three difficult situations to deal with in words that may be construed as locations: aliasing, in which two or more names are used to designate the exact geographic area, like "Mumbai" and "Bombay"; ambiguity between geo and non-geo, in which a word might refer to either a physical place or another type of thing; and geographic name ambiguity, in which different places might be having the same name. For example, the word "Aurangabad" is the name of two cities in India, making it difficult for disambiguation algorithms to connect the right place with it.

3. Proposed System

In MapNews, rather than just one new article, the system focuses on the geographic emphasis of a group of news stories on a similar theme, and this is accomplished using a document clustering method. Conversion of the document to feature vector [36] representation is the commonly used technique in document clustering, which is accomplished by using TF-IDF [37] method. Then using the cosine similarity measure [38], which is a simple distance function, these feature vectors are clustered. Two feature vectors are most likely referring to the same news story if they are within a distance E of each other. We can use the vector space embedding [40] or indexed [4] techniques to find similarities.

Every day, MapNews collects the most recent news from a variety of sources and analyses hundreds of new stories. Hence, the quick processing of individual articles and scalability of the system was the essential requirements in creating MapNews' architecture. Furthermore, it must also be resilient to failure by providing the latest news as soon as possible whenever they get published online.

MapNews works in a distributed computing cluster to facilitate the well-organized distributed processing of articles. For this purpose, we split news gathering and analysis phases into multiple modules and operate them individually on distinct processing nodes. News articles are processed in a computational pipeline by a series of these modules, as shown in Figure 1. A single article may be handled by multiple distinct processing nodes in the system since each module may run on a different node. Furthermore, it is possible that we can run several instances of each module on one or more nodes at the same time. As a result, for managing the volume of information, we may run as many copies of modules as needed. Each module gets data and delivers it to a synchronization point, which is a transactional database system. Additionally, the system can manage the increased system demand by replicating the database system across several nodes. These tasks are carried out using the PostgreSQL database package.

To coordinate the entire system, we built a unique master controller module apart from the individual processing modules. The news articles are assigned for processing to other modules (slave nodes) by the controller in the system. As an article moves through the connected slaves and the design, its progress is monitored by the controller using its database tables. For allocation tasks and notifying success or failure, an essential communication protocol is used between the master and slaves to send numerous control signals. When a slave module is created, it initiates a handshake once it connects to the master and indicates its availability and the task it will perform. Then slave will be allocated several articles to process by the master, and after processing slave will send a success or failure notification to the master. If the slave fails to answer to the master within a certain amount of time, the master thinks the slave has failed somehow.

3.1. News Crawling

Many media news organizations publish their stories and opinions on the Internet for a global audience. It can be challenging to automate the gathering and standardization of enormous news stories from such a broad range of sources. Many news sources over the Internet are excellent news sources, but many such sources are far from consistent. Major newspaper articles are typically internally consistent and well-formatted, but news blog websites may be questionable. Furthermore, the majority of newspapers, with a few notable exceptions, have a local emphasis and therefore concentrate on stories that are exclusive to a certain region. Consequently, rather than focusing on the most significant or most frequently distributed news sources, collecting news articles from news providers across the globe are equally important.

To get around these issues, MapNews focuses on a massive number of Really Simple Syndication (RSS) feeds as each published news item must be having a brief description, a news title, and a web URL. Optional publishing date is also available in RSS 2.0, which aids in determining the "freshness" of articles.

MapNews follows below steps in order to crawl the news articles from news websites.

1. Collecting the RSS feeds from online news sources.

2. Periodically scanning the RSS feeds for new stories.
3. Downloading and processing the new stories.
4. Extracting the content from the downloaded webpage.
5. Removing any unnecessary text or markup from the content.
6. Clustering the articles based on their content.
7. Removing any problematic sources or clusters by ranking clusters based on the number of distinct
8. News sources and tracking how well stories from each source cluster with stories from other sources.
9. Handling character encodings and removing any articles that are unable to be converted to a standard encoding.

By following above mentioned steps, MapNews's news crawling module collects and standardizes news articles from a variety of online news sources across the globe. It does this by using Really Simple Syndication (RSS) feeds, which provide a brief description, a news title, and a web URL for each published news item, as well as an optional publishing date. MapNews keeps track of current RSS newsfeeds from various online news providers and downloads and processes the latest articles when these feeds are checked for new ones on a regular basis. The system retrieves the article data from the downloaded webpage and uses strict content extraction rules to exclude unfitting articles and avoid extracting unnecessary words. It also removes articles that are unable to match the standard character encoding after conversion.

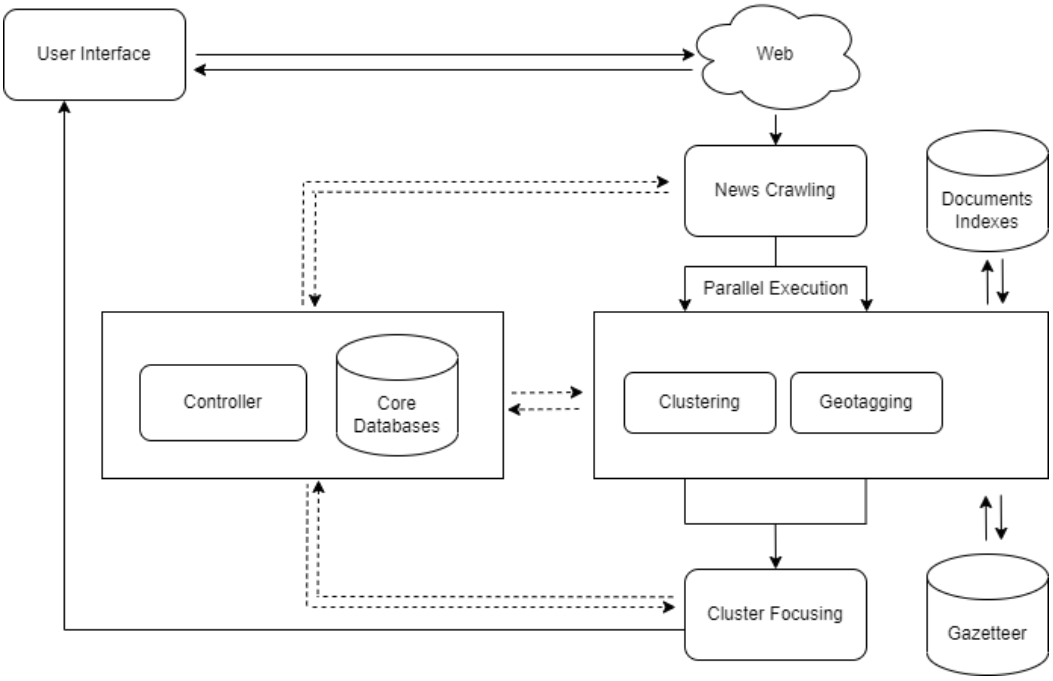

Figure 1. MapNews system architecture.

3.2. Geotagging

MapNews must find and extract the geographic information from the latest news story once it is uploaded to the framework. This method, known as geotagging, connects article text with the implied geolocation, allowing for the geographical study of the news [18]. We divide the geotagging module in MapNews into four phases as discussed below:

3.2.1. Entity Feature Vector Extraction

Extraction of "interesting" sentences is the first computation step, which might contain possible references to places and objects. All of these phrases make up the Entity Feature Vector (EFV) of the article. For acquiring the entity feature vector of an article, many researchers have discussed numerous approaches, such as TF-IDF [37]. However, for Named-Entity Recognition (NER) [5], we used Spacy [39]. Finding text words that correspond to object categories like LOCATION, ORGANIZATION, and PERSON is the goal of NER. The location-tagged sentences are probably the actual location names and saved as geographic features. In contrast, non-geographic attributes are assigned to the organization, and people tagged phrases in the entity feature vector as shown in Figure 2. The LingPipe toolkit's [26] NE tagger we used to be trained on the famous Brown [16] corpus and the MUC-6 conference news data.

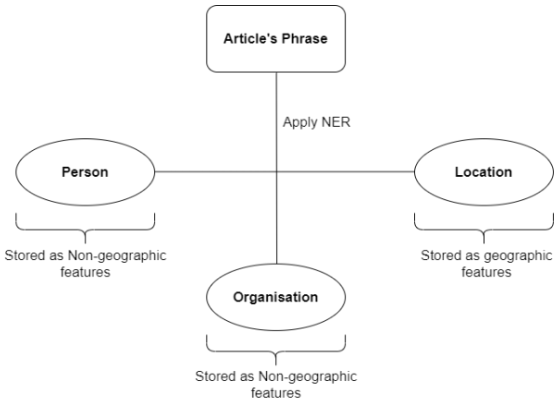

Figure 2. Phrase identification.

3.2.2. Assigning Gazetteer Record

MapNews uses a gazetteer or database to detect geographic characteristics in the entity feature vector once the entity features vector have been extracted. Gazetteer is a GeoNames database [17] used by MapNews to identify such geographic features, names of real locations in the entity feature vector. GeoNames database gazetteer used by MapNews contains more than 8 million location names and 6.5 million geographic locations across the world since GeoNames collects geographic data from over 100 sources. Every record in the gazetteer contains valuable information for geotagging, such as latitude and longitude and other names in different languages.

3.2.3. Geographic Name Disambiguation

Every geographic characteristic f ϵ EFV in MapNews is linked to a collection of corresponding places from the gazetteer, marked by the L (f). But certain attributes will be connected with numerous records, a consequence of ambiguous geographic names (i.e., |L(f)| > 1). As a result, MapNews reaches the stage of toponym resolution, also known as geographic name disambiguation [23]. This segment makes use of a number of heuristic filters to attempt to solve confusing references with the aid of deciding on a probable group of mappings for every reference, primarily depending on how someone could understand the news. These criteria are based on the premise that the places discussed in the text are linked concerning hierarchical confinement, document distance, and geographic distance.

3.2.4. Determining Geographic Focus

In this step, our tagger ranks the geo-references according to their importance to the article's geographic emphasis, separating those significant to the content from those referenced indirectly. The frequency of recurrence in the body text is one fundamental measure of importance. In addition, we have found that essential geo-references appear at the beginning of the text or title of a classic media story complying with a high geographic focus. To balance these two characteristics, we calculated a gradually declining weight of the geo-reference frequencies by giving geo-reference g's ranking higher importance whenever g appears closer to the start of the media story.

4. Applying Clustering Algorithm

The news articles that reflect the same news story should be aggregated using a clustering algorithm into a story cluster. Roughly, a news story is characterized by the content of the story as well as the duration it lasts, which means many of the exact essential keywords should appear in articles in a similar cluster, as well as publication dates that are near together.

Moreover, newly added articles should be grouped rapidly, such that readers may see the newly added news items right away. The news library would have to be clustered anew for each new article retrieved, resulting in severe performance penalties on heavy news days. Alternatively, we employ an online or incremental clustering technique that utilizes preexisting clusters and computes much faster. We also leverage the aforesaid time-based restriction and various optimizations to analyze hundreds of articles every day in real-time.

We employ the document vector space model [15] which is frequently utilized in information retrieval and text mining. A text document of this model in a d-dimensional space is represented as a term feature vector. The range of dissimilar words from each text in the dataset is denoted by d. The frequency of each expression in the text is represented by every term feature vector's element, which is calculated using a term weight algorithm. Whenever we add or remove an article from the system, d will vary, which should be recorded into the clustering method. Additionally, O(d) distance calculations may be excessively costly since d may have values of 100,000 or more and thus making the vector space frequently high-dimensional. Although to speed up distance calculations and achieve excellent performance, we utilize sparsity of the term feature vectors.

4.1. Data Preprocessing

We begin data preprocessing by eliminating punctuation and other non-essential symbols from a news article's content before clustering it. The TF-IDF value for every phrase in the news item is then used to derive the term feature vector of the news item. The computed value highlights words that often appear in a single text but are not commonly found throughout the corpus. For a term ti, the TF-IDF value is given below in Equation (1):

$$TF - IDF_{i,j} = \frac{n_{i,j}}{n_j}.\ log\frac{|D|}{o_i} \quad (1)$$

Where in the number of words in *dj* is represented by *nj*,

the number of times ti occurs in *dj* is represented by *ni,j*, and the articles number in D that include ti is represented by *Oi*.

4.2. Clustering Approach

The clustering algorithm we use is a variant of leader-follower clustering that allows for online clustering in both the term vector space (based on the term centroids) [15] and the temporal dimension (based on the time centroids). This means that the algorithm can identify groups of articles that are similar in terms of the importance or relevance of particular terms (as reflected by the term centroids) and the time at which the news articles were published (as reflected by the time centroids).

Maintaining a term centroid and time centroid for each cluster allows the algorithm to track the means of the term feature vectors and the publication times of the articles in each cluster. This can be useful for characterizing the overall content and temporal distribution of the articles within each cluster, and for identifying patterns or trends in the data.

We search for a cluster with a distance of less than a specified cutoff distance E between its time and term centroids and a. whenever there is a need to cluster a new article 'a'. The 'a' is added to the nearest cluster if more than one candidate cluster exists and then updates the cluster's centroids. If this is not the case, then a new cluster is formed with just 'a'.

With the cosine similarity measure [38] we may determine how far apart a candidate cluster and new article are in terms of its terms. For cluster c and article, a, the term cosine similarity metric is described as in Equation (2):

$$\delta(a,c) = \frac{\overrightarrow{TFV_a} \cdot \overrightarrow{TFV_c}}{||\overrightarrow{TFV_a}|| \quad ||\overrightarrow{TFV_c}||} \tag{2}$$

The term feature vector k is denoted by the TFV.

To incorporate the temporal dimension in clustering, we apply a Gaussian attenuator to the cosine distance measure, which gives more weight to clusters whose time centroids are close to the article's publication time. This can be done by modifying the distance measure used in the clustering algorithm.

For example, suppose we have a set of articles, each with a publication time and a set of features, and we want to cluster them based on the similarity of their features. We can use the cosine distance measure to compute the similarity between the articles based on their feature vectors.

The reason for using a Gaussian function in this case is that it provides a smooth, continuous transition between points that are close in time and points that are far apart in time. This allows for a more fine-grained control over the degree of attenuation applied to the distance measure, as the attenuation can vary smoothly based on the time difference between the points. Using a Gaussian attenuator can be useful in situations where we want to favor clusters of points that are close in time, as it allows us to give more weight to clusters whose time centroids are close to the article's publication time. This can be useful for finding patterns or trends in the data that are related to the time at which the articles were published.

The Gaussian parameter considers the time difference between the publishing time of the new article and cluster's time centroid. We present our new distance formula in Equation (3) as:

$$\dot{\delta}(a,c) = \delta(a,c) \cdot e^{\frac{-(T_a - T_c)^2}{2(2.2)^2}} \tag{3}$$

The publication time of *a* is represent by *Ta*, whereas the time centroid of *c* is indicated by *Tc*.

Cluster centroids are kept in an inverted index to improve performance, pointing to all clusters with a positive t value for each term t. The inverted index minimizes the distance calculations that must be performed during clustering. There are no distances computed to new articles ‘a’ that are grouped unless they have values other than zero. We also keep a running roster of recently active clusters with centroids as an additional measure of efficiency. If a cluster is on the active list, it will be considered for a new article, but not otherwise. Our distance function values become negligible after a few days, so we eliminate clusters from consideration. When these improvements in our method are combined, clustering articles may need fewer distance calculations.

5. Cluster Focus

We want to determine the cluster emphasis of articles in the same manner as we calculated the geographic focus of individual documents while geotagging. This means figuring out which places mentioned in news clusters are really essentials to the articles and which ones serve as references. The user interface will be shown using the locations identified during cluster focus calculation. However, despite the fact that we simply used word similarity to create clusters, this grouping makes sure that various variants of the same article are gathered together, which should also ensure that the included geo references are grouped. In order to accurately compute cluster focus, we combine the geotagging data from each article in the cluster. To be more specific, we assign a score to each item in cluster C that includes location l and mentions it at least once. Auto encoders are employed to form complex nonlinear encodings of the target expressions [28].

This process may be hampered by items that have been geotagged improperly, introducing location inaccuracies. It is possible to correct specific article inaccuracies at the cluster level by utilizing geotagging confidence values from enclosed articles and aggregated entity information. We can significantly enhance the quality of our cluster focus calculation if we establish

acceptable assumptions about story clusters and utilize certain information found while analyzing each document separately. Articles from the same cluster are assumed to relate to the same thing when two or more things identified in those articles share a name. A cluster of articles that all reference the entity "Ambedkar" refers to the same "Ambedkar", whether it be a person, a place, an organization, or another entity type. Because each article in a cluster is in the same category, we anticipate this assumption applying to story clusters. It's unusual for an article to reference two different places with the same name. It is more likely that a person or organization in the article will have the same name as a place in the article, although this is indeed an uncommon occurrence. This means that any differences between articles in a collection will most likely be due to geotagger inaccuracies.

For entities with inconsistently tagged tags, we use weighted voting to determine which ones are the right ones by utilizing our assumption. When one article in the cluster talks about e, every article votes on how they see e. The system gave votes for certain entities a boost in articles labeled with a greater level of confidence. It's possible that several articles will refer to "Mr. Ambedkar," implying that "Ambedkar" is widely understood as a personal name, in which case these articles will throw significant votes in favor of the understanding of "Ambedkar". However, if an article states "Ambedkar" and tags it as a place, it would get less vote for this understanding. Counting votes leads us to conclude that Ambedkar should be excluded from the cluster since it is a given name.

To make inconsistent locations more manageable, this concept may be used to letting readers vote on how location elements are interpreted in different articles. It's reasonable to believe that news reports on College Park in Mumbai will include references to the location of the park as well as mentions of "College Park," which is in Delhi. Because the first set of articles contains organizations that qualify as "College Park," they cast more votes in support of the location of College Park in Mumbai, and the accumulation of votes will also result in the positioning of College Park in Mumbai. There is no longer any competition for the cluster focus in Delhi's College Park. Discrepancies in entity interpretations are resolved by collecting C articles most often cited locations, and cluster C's focus is then determined. No matter how big or small a cluster is, we've found that the techniques outlined above work well for establishing cluster focus.

6. Map Interface

MapNews' user interface included as many geographical and non-geographical details about recent events as feasible. News articles are put on the map and shown in a window that we may use to search news by geographic area. Pan and zoom features are used by users to engage with MapNews and obtain more news articles.

Whenever a user pans or zooms on the map, new stories are retrieved for the viewing window and kept there so that it is always full of articles regardless of where the user is or how far they have zoomed in.

A specific map view attempts to produce an overview of the media articles in the display, integrating the significance of the articles with their geographical distribution throughout the globe. Users may zoom in or out to see additional articles about aspecific geographic area if they're just interested in aportion of the map's coverage.

MapNews uses Microsoft Virtual Earth's mapping API to show news in a web browser [21]. An example of MapNews' user interface can be shown in Figure 3, where several stories from India can be seen. The system places a marking on the map for each news cluster's geographic focus. The presence of a marker provides extra information about the news article that the marker is associated with. The story's main window will be opened on right hand side along with the themes (e.g., Business, Politics, and Health). Furthermore, more big markers will be used for more critical stories (those having articles from a variety of publications) than for less significant events.

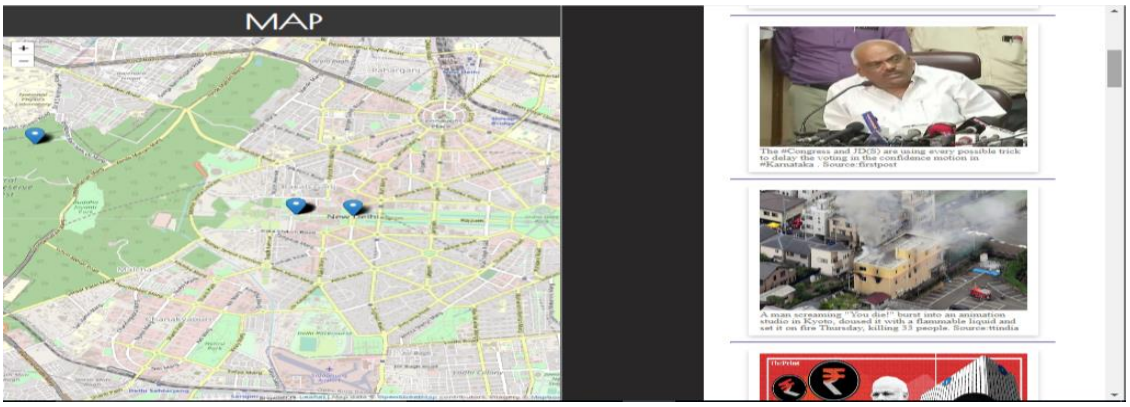

Figure 3. MapNews interface view, representing news clustered on New Delhi.

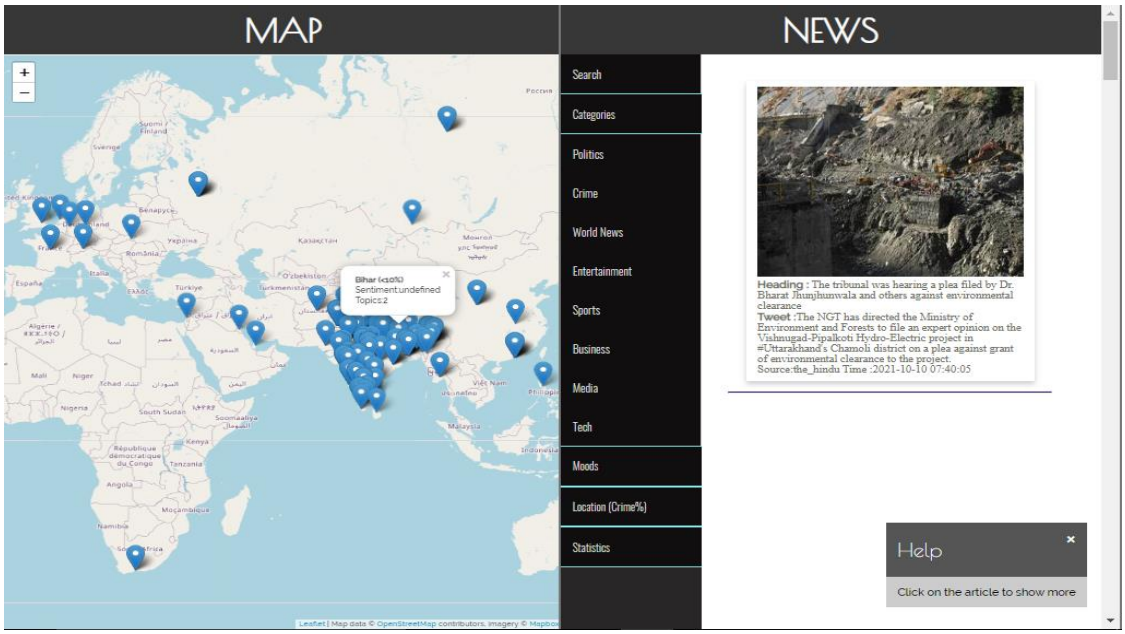

Figure 4. MapNews interface representing news articles in category mode.

When the mouse cursor hovers over a story marker, a tiny info-bubble appears, containing an overall overview of the article's content. A smaller map is also available in MapNews, which depicts the geographic scope of the chosen article. Using this minimap, readers may quickly see where the focus of the content is geographical without leaving the main map. Figure 4 shows the category mode in MapNews. Users may rapidly grasp the most essential or noteworthy subjects in the news by using the category mode instead of hovering over markers. Since categories take up more screen area than markers, the number of articles displayed on a map is limited.

7. Performance Evaluation

In this study, we evaluated the accuracy and usability of the MapNews using the Adjusted Rand Index (ARI) [42] and usability testing. The ARI was used to measure the alignment of the news aggregator's clusters with the known locations of the news articles, while the usability testing was used to assess the ease of use and effectiveness of the system. The ARI a widely accepted metric for comparing clustering. We also computed a contingency table [9] to visualize the overlap between the news aggregator's clusters and the known locations.
We have used a news dataset which includes news articles and their known locations to compare with the MapNews results for 50 news articles to compute the contingency table as shown in below Table 1.

Table 1. Contingency table.

	Location A	**Location B**	**Location C**
Cluster1	50	0	0
Cluster2	0	44	6
Cluster3	0	5	45

To solve the equation for the Adjusted Rand Index (ARI) using the contingency table provided above, we can use the formula given in Equation (4).

$$ARI = \frac{\frac{a+d}{a+b+c+d} - [(a+b)*(a+c)+(c+d)*(b+d)}{[(a+b+c+d)*(a+b+c+d-1)]} \quad (4)$$

Where,

- a: Number of news articles that are in the same cluster in both the MapNews's results and the known locations of our news dataset.
- b: Number of news articles that are in different clusters in the MapNews's results but the same cluster in the known locations of news dataset.
- c: Number of news articles that are in the same cluster in the MapNews's results but different clusters in the known locations.
- d: Number of news articles that are in different clusters in both the MapNews's results and the known locations news dataset.

This contingency table shows that the MapNews's clusters are highly aligned with the known locations, as indicated by the high ARI score of 0.8912.

In particular, all of the news articles in Location A are correctly assigned to Cluster 1, and most of the news articles in Locations B and C are correctly assigned to Clusters 2 and 3, respectively. There are a few instances where news articles are misclassified, such as some articles in Location B being placed in Cluster 3, but these instances are relatively rare and do not significantly impact the overall accuracy of the MapNews.

We also conducted usability testing for a news aggregator in order to evaluate the ease of use and effectiveness of the system. We used a combination of lab-based and remote testing methods, recruiting a total of 50 participants. We collected data on a variety of metrics, including task completion time, accuracy, and user satisfaction. The following Table 2 summarizes

the results of the usability testing:

Table 2. Usability testing.

Metric	Mean	Standard Deviation
Task Completion time(seconds)	60	15
Accuracy (%)	95	3
User Satisfaction (1-5 scale)	4.5	0.5

The mean task completion time of 60 seconds was within the expected range and did not vary significantly between the lab-based and remote testing methods. The average accuracy rate of 95% was also good, with most participants able to complete tasks accurately.

User satisfaction was generally high, with a mean score of 4.5 out of 5. However, there was some variability in the satisfaction scores, with a standard deviation of 0.5. This suggests that while most participants were satisfied with the news aggregator, there were some users who had a less positive experience.

In addition to the usability testing, we also conducted a survey with open-ended questions to gather more detailed feedback from participants. Some common themes that emerged from the survey responses included:

- The need for more personalized recommendations and a wider range of news sources: Several participants mentioned that they would like to see more personalized recommendations based on their interests, as well as a wider range of news sources to choose from.
- The importance of reliability and credibility: Many participants emphasized the importance of ensuring that the news aggregator only includes reliable and credible sources.
- The value of customization and control: A number of participants expressed a desire for more customization options, such as the ability to create their own news feeds or to customize the layout of the news aggregator.

Overall, the results of the usability testing and user feedback indicate that there is room for improvement in the MapNews, particularly in terms of navigation and search functionality. However, the system received generally positive ratings from users and there is a clear demand for personalized recommendations and a wider range of news sources.

8. Conclusions

MapNews has room for development in many areas. MapNews tends to favor regions where news articles are often published, so a broader range of perspectives is needed. Additionally, the system currently only analyzes articles published in English, so we may consider adding news sources and articles in various languages to enhance the design.

MapNews' geotagger could use more semantic indications from the text, like landmarks and rivers, to help with accurate geotagging. Instead of just relying on textual keywords, we will also consider using geography to better cluster news stories in the future. If clustering can be used to identify the geographic extent of the news source, it could be used to enhance local news coverage and geotagging. The map will also be enhanced with other content like photos, audio, and video clips, so we are looking at techniques to determine which image best represents the whole cluster of news articles.

The findings of the usability testing are encouraging in terms of the system's overall design, usability, and public acceptability. The layout, structure, and coherence all seem to meet current standards. However, user comments generally indicated that there are still challenges that need to be addressed in order to provide a better overall experience for all users.

References

[1] Allcott H. and Gentzkow M., "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211-36, 2017.

[2] Aniche M., Treude C., Steinmacher I., Wiese I., Pinto G., Storey M., Gerosa M., "How Modern News Aggregators Help Development Communities Shape and Share Knowledge," *in Proceedings of the International Conference on Software Engineering*, Gothenburg, pp. 499-510, 2018.

[3] Athey S., Mobius M., and Pal J., "The Impact of Aggregators on Internet News Consumption," National Bureau of Economic Research, Working Paper No. w28746, 2021.

[4] Bayardo R., Ma Y., and Srikant R., "Scaling up All Pairs Similarity Search," *in Proceedings of the 6th International World Wide Web Conference*, Banff Alberta, pp. 131-140, 2007.

[5] Belwal R., Rai S., and Gupta A., "Text Summarization Using Topic-Based Vector Space Model and Semantic Measure," *Information Processing and Management*, vol. 58, no. 3, pp. 102536, 2021.

[6] Burger J., Henderson J., and Morgan W., "Statistical Named Entity Recognizer Adaptation, Available: https://aclanthology.org/W02-2003, Last Visited, 2022.

[7] Buyukokkten O., Cho J., Garcia-Molina H., Gravano L., and Shivakumar N., "Exploiting Geographical Location Information of Web Pages," *in Proceedings of the WebDB (Informal Proceedings)*, Link oping, pp. 91-96, 1999.

[8] Calzada J. and Gil R., "What Do News Aggregators Do? Evidence from Google News in Spain and Germany," *Marketing Science*, vol. 39, no. 1, pp. 134-167, 2019.

[9] Carrizosa E., Guerrero V., and Romero Morales

D., "On Mathematical Optimization for Clustering Categories in Contingency Tables," *Advances in Data Analysis and Classification*, pp. 1-23, 2022.

[10] Chen Y., Suel T., and Markowetz A., "Efficient Query Processing in Geographic Web Search Engines," *in Proceedings of the ACM SIGMOD International Conference on Management of Data*, Chicago, pp. 277-288, 2006.

[11] Cucerzan S. and Yarowsky D., "Language Independent NER Using a Unified Model of Internal and Contextual Evidence," *in Proceedings of the 6th Conference on Natural Language Learning*, Stroudsburg, 2002.

[12] Das A., Datar M, Garg A., and Rajaram S., "Google News Personalization: Scalable Online Collaborative Filtering," *in Proceedings of the 6th International World Wide Web Conference*, Banff, pp. 271-280, 2007.

[13] Ding J., Gravano L., and Shivakumar N., "Computing Geographical Scopes of Web Resources," *in Proceedings of the 26th VLDB Conference*, Cairo, 2022.

[14] Dos Santos C. and Guimarães V., "Boosting Named Entity Recognition with Neural Character Embeddings," arXiv preprint arXiv:1505.05008, pp. 25-33, 2015. Available: https://arxiv.org/abs/1505.05008v2, Last Visited, 2021.

[15] Duda R., Hart P., and Stork D., *Pattern Classification*, Wiley-Interscience Publication, 2006.

[16] Francis W., "A Standard Corpus of Edited Present-Day American English," *College English*, vol. 26, no. 4, pp. 267-273, 1965.

[17] GeoNames., https://www.geonames.org/ Last Visited, 2021.

[18] George L. and Hogendorn C., "Local News Online: Aggregators, Geo-Targeting and the Market for Local News*," *Journal of Industrial Economics*, vol. 68, no. 4, pp. 780-818, 2020.

[19] Google News, "Google News," https://news.google.com/topstories?hl=en-IN&gl=IN&ceid=IN:en Last Visited, 2021.

[20] Kilimci Z. and Omurca S., "Enhancement of the Heuristic Optimization Based on Extended Space Forests Using Classifier Ensembles," *The International Arab Journal of Information Technology*, vol. 17, no. 2, pp. 188-195, 2020.

[21] Kloog I., Kaufman L., and De Hoogh K., "Using Open Street Map Data in Environmental Exposure Assessment Studies: Eastern Massachusetts, Bern Region, and South Israel as a Case Study," *International Journal of Environmental Research and Public Health*, vol. 15, no. 11, pp. 2443, 2018.

[22] Langseth A., "Use of Spatial Information in News Recommenders," [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3024702, Last Visited, 2022.

[23] Leidner J., "Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names," *ACM SIGIR Forum*, vol. 41, no. 2, pp. 124-126, 2008.

[24] Li S., "Replacement or Complement: A Niche Analysis of Yahoo News, Television News, and Electronic News," *Telematics and Informatics*, vol. 34, no. 4, pp. 261-273, 2017.

[25] Li Y., Shetty P., Liu L., Zhang C., and Song L., "BERTifying the Hidden Markov Model for Multi Source Weakly Supervised Named Entity Recognition," *in Proceedings of the Fifty-ninetieth Annual Meeting of the Association for Computational Linguistics and the eleventh International Joint Conference on Natural Language Processing*, Bangkok, pp. 6178-6190, 2021.

[26] LingPipe, http://www.alias-i.com/lingpipe/, Last Visited, 2021.

[27] McCurley K., "Geospatial mapping and navigation of the Web," *in Proceedings of The 10th International Conference on World Wide Web*, Hong Kong, pp. 221-229, 2001.

[28] McNamee P. and Mayfield J., "Entity Extraction without Language-Specific Resources," *in Proceedings of the 6th Conference on Natural Language Learning*, Sanya, pp. 1-4, 2002.

[29] Microsoft Live News., "Recent News-Stories," https://news.microsoft.com/recent-news/ Last Visited, 2021.

[30] Molina-Villegas A., Muñiz-Sanchez V., Arreola-Trapala J., and Alcántara F., "Geographic Named Entity Recognition and Disambiguation in Mexican News using Word Embeddings," *Expert Systems with Applications*, vol. 176, pp. 114855, 2021.

[31] Patrick J., Whitelaw C., and Munro R., "SLINERC: The Sydney Language-Independent Named Entity Recogniser and Classifier," *in Proceedings of the 6th Conference on Natural Language Learning*, Taipei, 2002.

[32] Pérez Sechi C. and Pérez Sechi C., "Leveraging Entities Knowledge to Bypass the Cold-Start Recommender Problem on Microsoft News Dataset," Máster en Minería de Datos e Inteligencia de Negocios, 2021.

[33] Phelan O., McCarthy K., Bennett M., and Smyth B., "Terms of a Feather: Content-based News Recommendation and Discovery Using Twitter," *in Proceedings of the Advances in Information Retrieval-33rd European Conference on IR Research*, Dublin, pp. 448-459, 2011.

[34] Ravin Y., Watson T., and Wacholder N., Extracting Names from Natural-Language Text, Citeseer, 1997. [Online]. Available:

http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.6337, Last Visited, 2021.
[35] Reddit News, "Reddit-Dive into Anything," https://www.reddit.com/ Last Visited, 2021.
[36] Salton G., Wong A., and Yang C., "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, 1975.
[37] Salton G. and Buckley C., "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513-523, 1988.
[38] Steinbach M., Karypis G., and Kumar V., "A Comparison of Document Clustering Techniques," 2000.
Available: http://conservancy.umn.edu/handle/11299/215421, Last Visited, 2021.
[39] Vasiliev Y., "Natural Language Processing with Python and spaCy: A Practical Introduction-Google Books," https://nostarch.com/NLPPython Last Visited, 2022.
[40] Wang S. and Koopman R., "Clustering Articles Based on Semantic Similarity," *Scientometrics*, vol. 111, no. 2, pp. 1017-1031, 2017.
[41] Yahoo News, "Yahoo News-Latest News and amp; Headlines," https://news.yahoo.com/ Last Visited, 2021.
[42] Zhang S. and Wong H., "ARImp: A Generalized Adjusted Rand Index for Cluster Ensembles," *in Proceedings of the 20th International Conference on Pattern Recognition*, Istanbul, pp. 778-781, 2010.
[43] Zhou G., and Su J., "Named Entity Recognition Using an HMM-Based Chunk Tagger," *in Proceedings of the Fortieth Annual Meeting on Association for Computational Linguistic*, Philadelphia, pp. 473-480, 2002.

Jeelani Ahmed obtained a Bachelor of Engineering and a Master of Technology in Computer Science from Visvesvaraya Technical University in Belagavi, India in 2012 and 2015, respectively. He is currently working towards a PhD in Computer Science at Maulana Azad National Urdu University in Hyderabad, India, which he has been doing since 2017. His research interests include Big Data Analytics, Semantic Web, Network Security, and Cloud Security. In addition to his academic pursuits, Jeelani Ahmed has six years of teaching experience and five years of research experience. He has presented and published numerous research papers at national and international conferences and in academic journals.

Muqeem Ahmed is an Assistant Professor at the Department of Computer Science and Information Technology, Maulana Azad National Urdu University, Hyderabad, India. He received his doctoral degree in computer science from Jamia Millia Islamia in New Delhi, India. He has more than 14 years of experience in teaching, research, and project supervision. Throughout his career, he has supervised students in interdisciplinary research and industrial projects and has published numerous research papers in national and international journals. In addition to his academic work, Dr. Muqeem Ahmed is also a member of the editorial boards and reviewer panels of various journals. His main areas of research include semantic web applications, distributed databases, machine learning, and big data analytics.