Abstract— In this quick moving world, electronic media assumes significant part in individuals' normal life. Electronic media conveys most recent news and exceptional data faster than conventional daily papers. A portion of the news advantages individuals others not. Individuals not able to sufficiently extra time to peruse whole news from daily papers or listen media like TV or radio. The propensity for news perusing has been changed with the headway of World Wide Web (WWW) from conventional model of news utilization through Newspaper membership to current time Internet. Presently a-day's PDA innovation conveys the web office to the individuals whenever. This makes individuals acquainted with the utilization of online news perusing and redesigns most recent data on the go. Online news perusing got to be well known as the web gives access to news articles from a huge number of sources the world over web. Media distributors post the news in the web gateway and individuals can discover them through some web index like google or direct media site. On the off chance that the news substance is rich in content, great picturization, sifting and rundown of most recent news, then get much consideration among the news perusers. In this paper, we analyzes the different models of Content based proposal and Collaborative suggestion. We join both these systems to make a crossover proposal framework as an answer for the issues of news suggestion. This mixture news suggestion method removes news from unique locales and presents to the perusers all the more viably. This builds the news perusing enthusiasm of individuals furthermore expands the hit rate of that site.

Keywords— News recommendation; Data mining; Web mining; K-Nearest Neighbour Algorithm; RSS Feed.

I. INTRODUCTION

Numerous real players, organizations, in this field are attempting to overcome existing issues to manufacture a successful news proposal framework which fulfills client. The accompanying are the current issues to be tended to assemble a successful news recommender framework.
Absence of Data: Obviously the greatest issue confronted by recommender frameworks is that they require a great deal of information to successfully make suggestions. A decent recommender framework chiefly needs news thing and after that it must catch and look at client information.

Evolving Data: Systems are typically one-sided towards the old and experience issues demonstrating new. Past conduct of clients is not a persuading device in light of the fact that the patterns are continually evolving.

Changing User Preferences: The issue here is that while today client have a specific expectation when scanning one site - tomorrow client may have totally diverse aim. The framework which takes client inclinations as a base for proposal, might inaccurately name clients.

Unusual Items: This sort of issue by and large happens on account of some whimsical feature or music. These sorts of things are hard to make proposals on, in light of the fact that the client response to them has a tendency to be various and eccentric.

Excessively Complex: It takes loads of variables to do even least complex proposal.

Information mining is all the more fittingly named learning mining from information. The general objective of the information mining methodology is to concentrate data from an information set and change it into a reasonable structure for further utilize.

Web mining is the utilization of information mining systems to naturally find and concentrate data from Web records and administrations. At scale boundless, we concentrate on the last one removing quality from site pages and different archives found on the web. While inquiry is the greatest web excavator, the web mining results are profitable. Web mining may be separated into three classes as Web substance mining, Web structure mining and Web utilization mining.

Web substance mining: Discovering valuable data from substance of Web pages. Web substance is extremely rich comprising of printed, picture, sound, feature and metadata and also hyperlinks.

Web utilization mining: Making feeling of Web clients' conduct. This discovers client propensities which can help in rearranging a Web website so that high caliber of administration may be given. Existing instruments report the quantity of hits of Web pages and where the hits originated from. Albeit helpful, the data is not adequate to learn client conduct. Devices giving further examination of such data are helpful.
Web structure mining: Discovering the connection structure or model fundamental the Web. The model is in view of the topology of the hyperlinks. This can help in finding closeness between destinations or in finding power locales for a specific point or discipline or in finding review or study destinations that indicate numerous power destinations.

In this paper, we proposed a calculation for successful news suggestion framework. This calculation utilizes ideas of aggregate knowledge, handles the majority of the issues specified above and non-sufficiently specific to be connected.

II. RELATED WORKS ON NEWS EXTRACTION

F. Garcin, et. al. [1] introduced personalized news recommendation based on collaborative filtering system. There is abundance of news on the web an efficient news recommendation is a need of the hour. The authors compare three approaches for personalized news recommendation, they are: collaborative filtering at the level of news items, content-based system recommending items with similar topics, and a hybrid technique. He concluded that recommending items according to the topic profile of the current browsing session seems to give poor results.

J. Liu, et. al. [2] introduced personalized news recommendation based on click behaviour. Online news reading has become very popular as the web provides access to news articles from millions of sources around the world. A key challenge of news websites is to help users find the articles that are interesting to read. They developed a recommendation system for Google News. For users who are logged in and have explicitly enabled web history, the recommendation system builds profiles of users’ news interests based on their past click behaviour.

intelligence, aiming to find interesting news and summarize concise content for users. In their work, a keyword knowledge base was maintained and which provides a real-time update to reflect the general Web news topic information and the user’s interest preferences.

Jensen [5], proposed Bayesian framework for user interest prediction. In this work, the log analysis reveals that the click distributions of individual users are influenced by the local news trend. For example, Spanish users read more sports news during Euro Cup. Similar phenomena were also reported in a user study of the lifecycle of news interests. Based on these findings, author classifies user’s news interests into two parts: users’ genuine interests and the influence of local news trend. The user’s genuine interests originate from the personal characteristics of the user, such as gender, age, profession, etc. and are thus relatively stable over time. On the other hand, when deciding what to read, users are also influenced by the news trend in the location that they belong to. This kind of influence produces short-term effects and changes over time.

Morris and Hirst [5] introduced a concept of lexical chains to segment text. Later, lexical chains are used in many tasks, such as text retrieval and information extraction. The construction of lexical chains needs a thesaurus for determining relations between words. They constructed the lexical chains using the thesaurus-based word similarity and the word co-occurrence model. Two thesauruses, including WordNet and HowNet, were respectively used to compute word similarity in English and in Chinese. The notion of cohesion is a device for “sticking together” different parts (i.e., words, sentences, and paragraphs) of the text to function as a whole. Lexical cohesion occurs not only between two terms, but also among sequences of related words, called lexical chains.

D. Billsus and M. Pazzani [5] proposed a technique in which creating a model of the user’s preference from the user history is a form of classification learning. The training data of a
classification learner is divided into categories, e.g., the binary categories “items the user likes” and “items the user doesn’t like.” This is accomplished either through explicit feedback in which the user rates items via some interface for collecting feedback or implicitly by observing the user’s interactions with items. Users can click on the thumbs up or thumbs down buttons to indicate whether they like the program that is recommended.

D. Chakrabarti, et. al. [5] used the newsletters for the net-works are available as tabs on the top of the page to develop news recommendation system. The newsletter is presented as a list of articles, each with its title, news synopsis, links to original article and the locally cached page. Users can rate an article on a scale of 1 to 5. In the newsletter, the clusters are sorted on the score of the most relevant document in each cluster. The users can also recommend particular news articles to their friends through a recommendation button provided beside each news result.

X. Wu, et. al. [5] proposed a technique known as Keyword extraction, which is applied for document retrieval, web page retrieval, document clustering, summarization, text mining, and so on. By extracting appropriate keywords, we can easily choose which document to read to learn the relationship among documents. Keyword extraction from text data is a common tool used by search engines and indexes alike to quickly categorize and locate specific data based on explicitly or implicitly supplied keywords.

III. NEWS RECOMMENDATION SYSTEM

A. Content-Based Recommendation

A content-based recommender learns a profile of the user’s interests based on the features present in objects the user has rated. Content-based recommendation systems may be used in a variety of domains ranging from recommending web pages, news articles, restaurants, television programs, and items for sale. Although the details of various systems differ, content-based recommendation systems share in common a means for describing the items
that may be recommended, a means for creating a profile of the user that describes the types of items the user likes, and a means of comparing items to the user profile to determine what to recommend.

Content-based recommendation is an outgrowth and continuation of information filtering research. In a content-based system, the objects of interest are defined by their associated features. A content-based recommender learns a profile of the user’s interests based on the features present in objects the user has rated. The type of user profile derived by a content-based recommender depends on the learning method employed. Content-based user profiles are long term models and updated as more evidence about user preferences is observed.

B. Collaborative Recommendation

Collaborative recommendation is probably the most familiar, most widely implemented and most mature of the technologies. Collaborative recommender systems aggregate ratings or recommendations of objects, recognize commonalities between users on the basis of their ratings, and generate new recommendations based on inter-user comparisons. A typical user profile in a collaborative system consists of a vector of items and their ratings, continuously augmented as the user interacts with the system over time.

Some systems used time-based discounting of ratings to account for drift in user interests. In some cases, ratings may be binary (like/dislike) or real-valued indicating degree of preference. These systems can be memory based, comparing users against each other directly using correlation or other measures, or model-based, in which a model is derived from the historical rating data and used to make predictions. Model-based recommenders have used a variety of learning techniques including neural networks, latent semantic indexing, and Bayesian networks. The greatest strength of collaborative techniques is that they are completely independent of any machine-readable representation of the objects being recommended.

C. Recommendation Algorithm

Many recommendation methods can be directly applied to Web news personalization. However, Web news has several characteristics including dynamic content, changing interests, multiple interests, novelty, and so on, that make some approaches better suited than other approaches. Because collaborative methods suffer from the “latency” problem that needs some time to receive enough users’ feedback, content-based approaches are better
suited to the problem than collaborative approaches. In this work, we divide the recommendation news into three groups: previous news tracking, interesting topics, and novelty news. The proportion of the recommended news for each group is defined by the user. We use the $k$-nearest neighbour algorithm to track previously read news and find novelty news. The $k$-nearest neighbour algorithm identifies recently known stories that the user has read. It keeps tracking new stories that have the same event thread with recently read stories, and it finds novel news. After filtering out the non-news parts on the Web news page, each news article is converted to a TFIDF vector as follows:

$$\text{TFIDFI} = \frac{tfi \times \log \left( \frac{N}{ni} \right)}{\sqrt{\sum (tfi \times \log \left( \frac{N}{nj} \right))^2}}$$

where $tfi$ is the frequency of word $wi$ in the given Web page, $N$ is the number of documents in the corpus, and $ni$ is the number of documents in the corpus that contain word $wi$. The cosine measure is used to compute the similarity of two vectors. In this paper, we define three similarity thresholds: $t1$, $t2$, and $t3$ ($0 < t1 < t2 < t3 < 1$) to decide whether a news story is interesting, not interesting, or novel. We calculate the similarities of the coming news story with the most recently rated stories and search $k$ nearest neighbors. If one of the rated stories is closer than $t3$, the coming story is labeled as uninteresting (the user has known it). If the average of the $k$ similarities is less than $t1$, the story is labeled as novel; the smaller the average of the similarities is, the more novel the story is. If the average of the $k$ similarities is larger than $t2$, the story is labeled as interesting; the larger the average of the similarities is, the more the story is interesting.

In our recommendation system, $t1$, $t2$, and $t3$ are set to 0.05, 0.3, and 0.9, respectively, and $k$ is 20 based on our empirical observations. Each news story is represented as a feature-value vector, where features are the keywords selected from the news story, and feature values are the word frequencies. The user topic preference is also represented as a vector where keywords are selected from the total browsed stories.

**Proposition 1** Assume that user $u$ is independent to the news document $d$ given the news topic classification model $C = \{c1,c2,...,cn\}$, where $n$ is the number of news topic categories. The probability that document $d$ is recommended to user $u$ is computed as follows:
Proof: According to the conditional probability formula,

$$p(u \mid d) = p(u, d) \mid p(d),$$

and by the total probability theorem,

$$p(u, d) = \sum_{j=1}^{n} p(u, d \mid cj) p(cj)$$

For a given user, $p(u)$ is a constant value, so we can recommend $d$ to $u$ using the formula:

$$p(u \mid d) = \sum_{j=1}^{n} \frac{p(cj \mid u)p(u \mid d)}{p(cj)}$$

Then

$$p(u \mid d) = \sum_{j=1}^{n} \frac{p(cj \mid u)p(d \mid cj)p(cj)}{p(d)}$$

Since

$$p(u \mid cj)p(cj) = p(u)p(cj \mid u), \quad \text{And}$$

$$p(d \mid cj)p(d) = p(cj \mid d)p(cj)$$

$$p(u \mid d) = p(u) \sum_{j=1}^{n} \frac{p(cj \mid u)p(cj \mid d)}{p(cj)}$$

$$p(u \mid d) \propto \sum_{j=1}^{n} \frac{p(cj \mid u)p(cj \mid d)}{p(cj)}$$

Given the similarity thresholds $t_1, t_2, t_3$, the number of most nearest neighbours $k$, where $t_1$, $t_2$, $t_3$, and $k$ are decided by experiments, the algorithm outputs recommended stories according to the user’s rated histories in the most recent past.
We formalize the recommendation algorithm as follows.

1: **FOR** each upcoming news story **DO**
2: calculate the similarities of the news story with the user’s recently rated stories and get $k$ most nearest neighbours;
3: **IF** one of the $k$ similarities is larger than $t_3$
4: label the upcoming story as uninteresting;
5: **CONTINUE**;
6: **IF** the average of the $k$ similarities is larger than $t_2$
7: put the new story into the interesting queue;
8: **CONTINUE**;
9: **IF** the average of the $k$ similarities is less than $t_1$
10: put the new story into the novelty queue;
11: recommend the stories in the interesting queue in the descending order of the average similarity;
12: recommend the stories in the novelty queue in the ascending order of the average similarity;
13: recommend the remaining stories according to the probability calculated by formula.

**IV. PROPOSED HYBRID RECOMMENDER SYSTEM**

Hybrid recommender systems combine two or more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one. Most commonly, collaborative filtering is combined with some other technique in an attempt to avoid the ramp-up problem.

A weighted hybrid recommender is one in which the score of a recommended item is computed from the results of all of the available recommendation techniques present in the system. For example, the simplest combined hybrid would be a linear combination of recommendation scores. It initially gives collaborative and content-based recommenders equal weight, but gradually adjusts the weighting as predictions about user ratings are confirmed or disconfirmed. The benefit of a weighted hybrid is that all of the system’s capabilities are brought to bear on the recommendation process in a straightforward way and it is easy to perform post-hoc credit assignment and adjust the hybrid accordingly. However,
the implicit assumption in this technique is that the relative value of the different techniques is more or less uniform across the space of possible items.

In my proposed work, we combine the content based recommendation and collaborative recommendation into hybrid recommendation system. This system extracts the document and news information from various website to form the reliable information in our website (See Figure 1). We developed a News website which contains information from various news website and it reduces the user overload as well as hard time to find the correct information. It also increase the user interest and increase the site hits compare with many other website.

![Home page of hybrid recommendation system](image)

Fig. 1 Home page of hybrid recommendation system

V. RESULT AND DISCUSSION

In this research work, we developed hybrid recommendation system by combining content based recommendation and collaborative recommendation. Hybrid recommendation system performs effectively and displays news content extracted from their original news web site. Hybrid recommendation increases the user interest for news reading.

This hybrid recommendation system is developed using ASP.NET. The pages developed with ASP.NET are faster than Classic ASP. ASP.NET has better language support, a large set of user controls, XML-based components, and integrated user authentication. ASP.NET pages have the extension .aspx, and are normally written in VB (Visual Basic) or C# (C sharp).
To develop homepage of proposed hybrid recommendation system (See Figure 1), we extracted news from four different news websites such as Dhinakaran, Dhinamalar, The Hindu and The Times of India.

![Fig. 1 Content-Base Recommendation System](image1)

In Figure 2, the Content-Base recommendation system is depicted in which the news items are generated by extracting the original picture from sports websites and contents from local Tamil news site Dhinakaran.

![Fig. 2 Collaborative Recommendation System](image2)

The Figure 3 illustrates the news items generated using Collaborative recommendation system in which the news contents are extracted from a news website ‘The Times of India’.
In this method, we can customise the news items as desired by the intended user. In Table 1, the comparison between existing news recommendation systems are discussed and compared with our proposed hybrid news recommendation system.

### TABLE I

**COMPARISON BETWEEN EXISTING AND PROPOSED WORK**

<table>
<thead>
<tr>
<th>Author &amp; Method</th>
<th>Existing Methodologies</th>
<th>Proposed Hybrid methodologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jensen, proposed Bayesian framework for user interest prediction”</td>
<td>In this work, to understand the interest of the user is difficult since the genuine interests are based on personal characteristics, such as gender, age, profession, etc.</td>
<td>In this work, we combine both the collaborative and content based recommendation in to one hybrid method.</td>
</tr>
<tr>
<td>Morris and Hirst, proposed Lexical chains to segment text”</td>
<td>The word co-occurrence model was adopted to solve the problem. It was observed as difficult to compute the semantic relations between words not in the thesaurus.</td>
<td>The original contents are extracted from news website and updated news items which increase the user interest.</td>
</tr>
<tr>
<td>D. Billsus and M. Pazzani,</td>
<td>In this model constructed based</td>
<td>In our work, we generated news</td>
</tr>
</tbody>
</table>
proposed a model of the user’s preference which extracted from user history.

| on user interest on particular content types. It takes more time to understand users’ interest. | items based on user ratings. |

### VI. CONCLUSIONS

In this work, we focused on developing a system which recommends interesting news to users. This system recommends more interesting and important news after thorough analysis, since the nature of news reading makes news information filtering distinctive from information filtering in other domains. All existing recommender systems employ one or more of a handful of basic techniques such as content-based, collaborative and knowledge-based.

The content-based approach recommends items based on the profile which is built by analysing the content of articles that a user has read in the past. In collaborative filtering approach, the system uses known preferences of a group of users to make recommendation for other users. In this work, we developed a hybrid news recommendation system which uses RSS feed method to extract news from other popular news sites. This method combines Content-based approach and Collaborative filtering approach for an efficient hybrid news recommendation system. The performance and the way it recommends the news items are comparable with the relatively high performance recommendation systems existing now.

As a future work, we proposed to develop a more interactive model which can generates news items under particular label based on the user’s interest and hit rate.

### REFERENCES


