Classification of News Using Naïve Algorithm

Mamata Thakur[1], Priyanka Tamgadge[2], Pritam Thakur[3], Govind Rao Mettu[4]

Student[1,2,3], Faculty[4]

Department of Computer Engineering,
Pillai HOC College of Engineering and Technology, Rasayani, India

Abstract— **There are billions of web pages available on the Internet. Search Engines always have a challenge to find the best ranked list to the user's query from those huge numbers of pages. A lot of search results that corresponding to a user's query are not relevant to the user need. The incentive for this work originates from the need of retrieving useful web news pages from the Indian news websites corpus. News web pages contrast from other web pages; it is mainly vital to recognize web news accurately for precise classification. We will likely locate a simple yet efficient technique to mine news articles from web corpus. To accomplish this task, the automatic recognition method has been recommended for news web page classification that uses classification rules based on a combination of content, structure and uniform resource locator (URL) attributes. We gathered news web documents from 10 different news websites. We use Naïve Bayes algorithm to distinguish news articles from non-news articles examples advertisements, not related links.**

Keywords: - news web page classification, content attributes, structure attributes, URL attributes

I. INTRODUCTION

Classification shows an important part in various information retrieval tasks. The web is very diverse in nature, and no rules are there on how to build HTML pages and how to state the entire structure of the web pages. Thus automatic web page classification is an important task. Web page classification technique uses a variety of information to classify a target page. The vital notion for web page classification is the similarity measurement between web documents. Similarity analysis and classification can be done on attributes drawn from web documents.

Online newspaper websites have a standout amongst the most essential up to date information. Many websites provide day by day news in extremely different formats, and effective classification is required to get to and monitor this data in an automatic manner. Current web page classification methods use an assortment of data to classify a web page like the content of the page, structural information of the web page and the URL of the target page. Therefore a web page content, structure and URL are least expensive to achieve and significant sources for classification. In this paper, we randomly select ten newspaper websites as sources of information. In general news sites consist of thousands of web pages. These web pages are represented by vectors of vital features such as structure, URL, and content attributes.

A classification technique uses those features for the news web page recognition. The preferred pages are the news article page. In this way, the primary objective is to recognize which pages is the news pages and non- news pages screened out. Our method for web news page recognition depends on the collection of essential attributes and then uses a classification algorithm to classify web news pages.

Considered World Wide Web is the largest database in the Universe which is mostly understandable by human users and not by machines. It lacks the existence of a semantic structure which maintains interdependency of its components. Presently search on web is keyword based i.e. information is retrieved on the basis of text search of all available matching URL’s / hyperlinks. This may result in the presentation of irrelevant information to the user. In the current web, resources are accessible through hyperlinks to web content spread throughout the world.

II. RELATED WORK

Ranking search results is a fundamental problem in information retrieval. Most common approaches mainly focus on similarity of query and a page, as well as the overall page quality. However, with increasing popularity of search engines, the capturing of user behaviors insists to appear on the surface more. Much information such as links users click how long users spend on a page and the user's satisfaction degree from the relevance of the page could be estimated. It is actually kind of implicit feedback (i.e. the actions users take when interacting with the search engine), such kind of usage data could be used to improve the rankings. A lot of work has been done on the implicit measures of user preference in the field of IR (i.e. implicit feedback in IR).

Morita et al. in 1994:
One of the earliest evaluations of time aspects was presented by Morita et al. in 1994. Their experiments showed a positive correlation between user interest and the reading time of articles. In addition, they found a low correlation between reading time and the length and readability of an article.

Ding et al. in 2002:
Usage-based ranking Algorithm was presented by Ding et al. in 2002 for web Information Retrieval systems that applies time

spent on page against standard selection- frequency based ranking, i.e. the basic idea of rank score is calculated on the time users spend on reading the page and browsing the connected pages, the high- ranked pages may have a negative adjustment value if their positions couldn't match their actual usage, and the low-ranked pages may have a positive adjustment value if uses tend to dig them out from low positions.

Kellar et al. 2004:
According to the study of Kellar et al. 2004 focused on the relation between web search tasks and the time spent on reading results. Their results support the correlation and show that it is even stronger as the complexity of a given task increases.

Kritikopoulos et al:
Kritikopoulos et al. was studied method in for evaluating the quality of ranking algorithms. Success Index takes into account a user's click-through data, the result shows their method is better than explicit judgment.
A comparison study was appeared on between three methods of ranking in usage field. Those methods are Page Rank, Weighted Page Rank and HITS. All of those methods are focus on the structure of the page. The result of this comparison is HITS is the best.
In this research was presented a method based on a combination of click-through of pages by the users (event) and the summarization of documents. They used the advantage of implicit modeling is effectively improving the user model without extra effort of user. As result implicit feedback information improves the user modelling process.

Rekha et al. in 2011:
Another study was presented by Rekha et al. in 2011. This study was provided a new model to find a user's preferences from click-through behavior and using the exposed preferences to adapt the search engine's ranking function for improving search service. In this proposed model, the combination of viewed and stored document summaries is used. The results show that this combining improved the reliability of ranked- list than ever was.

Mukherjee et al. (2012):
Mukherjee et al. (2012), present a method to discover web knowledge for presenting web users with more personalized web content. Their method was collected usage data from different users and then finds the similarities between all pairs of users. Experimental results generate correct suggestions that retrieve relevant documents to the user.

Tuteja's study in 2013:
Tuteja's study in was based on user behaviors in order to enhance the weighted Page Rank Algorithm by considering a term Visits of Links (VOL) done by the end of 2013. This research idea presented as modifying the standard Weighted Page Rank algorithm by incorporating Visits of Links. Some usage behavior factors included in this research to VOL like:

- Time spent on web page corresponding to a link: The algorithm must assign more weight to the link if more time is spent by the users on the web page corresponding to that link. Most of the times, the time spent on the junk pages is very less as compared to relevant pages. So this factor will help in lowering the rank of junk pages to improve the classification.
- Most recent use of link: The link which is used most recently by users should have more priority than the link which has been not used so far. So most recent use of searched link can also be used to calculate the page rank.

The result shows that adding number of visits of links (VOL) to calculate the values of page rank holds to be more relevant results are retrieved first. In this way, it may help users to get the relevant information much quicker.

III. NEWS WEB PAGE CLASSIFICATION

Web page classification techniques use diverse information to classify a target web page: the content of the web page, web page URL and structure information on a web page. To classify the preferred news pages our approach identifies and explores common attributes that are commonly present in news websites. Maximum news websites are organized as
(i) Homepage that shows some headlines of all sections.
(ii) Numerous unit of pages that offer the headlines of diverse extents of interest like business, sports, entertainment, technology, politics etc. these different areas also contains some sub sections like national, international, market, cricket, football, science etc.
(iii) Pages that actually represent the news containing the title, author, related news link, date and body of the news. On the other hand web page URL is one of the most informative sources of information with respect to classification. URL of a web page is mainly content bearing, and it seems useful in making full usage of this resource. The aim of our approach is to classify correctly news pages disregarding the other pages. In this paper, we select URL attribute, structural attributes and content attribute for news web page classification.

IV. PROBLEM STATEMENT

With an enormous growth of the Internet it has become very difficult for the users to find relevant documents. In response to the user’s query, currently available search engines return a ranked list of documents along with their partial content. If the query is general, it is extremely difficult to identify the specific document which the user is interested in. The users are forced to sift through a long list of off-topic documents. Moreover, internal relationships among the documents in the search result are rarely presented and are left for the user.

The growth of the World Wide Web has enticed many researchers to attempt to devise various methodologies for organizing such a huge information source. Search engines were introduced to help find the relevant information on the web. However, search engines do not organize documents automatically they just retrieve related documents to a certain query issued by the user.

V. PROPOSED APPROACH

To classify news web pages correctly, we firstly choose some news websites after that select the important attributes from these news websites and create a dataset. Then naïve Bayes classifier is used to recognize news pages from non-news pages. The steps of the approach are as follows:

1. Some of online news sites are selected randomly.
2. By using Google search engine, we generate a dataset that contains content, URL and structural attributes.
3. We use Naïve Bayes classifier to identify the news pages from non-news pages.
4. We evaluate the performance of Naïve Bayes classifier in terms of precision by using WEKA tool.

A. *Attributes Selection*

URL Attributes: URLs are an extremely exquisite feature for learning. It is an important identification feature for web news; the URLs of news websites are often same structure. URL of news website contains both positive and negative attributes. For news web page identification positive attributes are more useful than negative attributes. Second level domain attributes and first level catalog attributes come under positive attributes list.

a) Positive attributes:

- Second level domain attributes: Similar sections of different news web pages share related structure attributes. For example URLs of subsections of news web pages like business, tech and sports also have second level domain attributes such as “business”, “tech”, and “sports”.
- First-level catalog attributes: URLs also contains first-level catalog attributes of news web pages such as “newspaper name” and news center. The First level catalog attributes provides a vital basis for the recognition of the news web page.

b) Negative attributes:

- Bbs
- Blog
- Video
- Ads
- Campaign

2) Content Attribute: After selecting 750 html pages of news web pages and 275 HTML pages of non-news web pages, which was selected randomly from 10 different news websites, we observed that the occurrence of the “news” keyword in a webpage is an essential attribute for the news web page recognition. News in the news websites are classified as politics, sports, business, etc. in every category, there are also subcategories for example, the subcategories, that occur in the sports category are cricket, golf, tennis, football, hockey, etc. In business, market, share, economy, etc. We selected some keywords as content attributes of a news web page: News Center, article source, author, related news, interconnected subject, connected link and count how many times the term news appear in the HTML page, date.

3) Structure Attributes: News web pages encompass rich structure information that can increase the accuracy of a classifier if correctly used. By analyzing the structure of different news web pages we observe that certain structure attributes contribute to news web page recognition, containing web page title and subtitle written as <title>, <Hn> tag and <div> tag that form up a webpage’s hierarchy. <title> tag of all news websites are similar and contains web page title or news center and website information like newspaper name. <div> tag contains the date and time feature of the webpage, which is necessary for news webpage recognition. The combined attributes of web news pages are as shown in Table I.

B. *Experimental Dataset*

Experimental dataset for the classification of news web pages described in this paper depends upon the attributes from 10 different Indian news websites. These websites are named as Times of India, Hindustan times, NDTV, Indian Express, The Hindu, The Pioneer, India Today, Deccan Herald, The Asian Age, The Telegraph. To build up the several corpus section of news web pages are reviewed such as sports, politics, entertainment, technology, business.

Attributes		
URL attributes	**Content attributes**	**Structure attributes**
Positive attributes: Second level domain: news, tech, sports, country name, Business, economy, politics,	News title Article source Author Top News Stories Latest News	Date and time feature in <div> tag Has news center and newspaper

science, Market, budget, gadgets, careers, world Games First- level catalog attributes: newspaper name **Negative attributes**: bbs, blog, video, ads, campaign,	Related news Related link Related subject Sum up the number of times the term news appear in the HTML page Date and Time	name in <title> tag <a> tag contain top news or related news

TABLE I. COMBINED ATTRIBUTES FROM TEN NEWS WEBSITES

C. *Learning algorithm*

We outline our learning algorithm in this section. We can usage any existing classification method to perform classification with these attributes. We classify using Naïve Bayes algorithm. This algorithm comes under probabilistic approach. For text classification applications and experiments generally Naïve Bayes classifier is used. The basic idea is to use Naïve Bayes classifier is the joint probability of words and categories to assess the probabilities of the classifications given a document.

Attribute selection from website: The Times of India, BBC News, News24, Daily Mail, Google News (India), NBC News, ESPN Cric Info, CNBC, Newsweek, TechCrunch, New York Magazine.

VI. SYSTEM FLOW

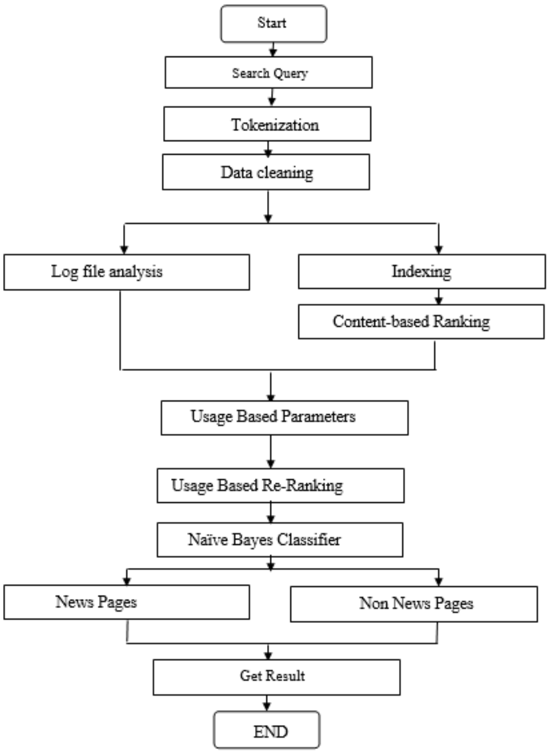


Figure 1. Work flow of system

1. Tokenization: This stage for breaking a stream of text up into words, and keeping the words in a list called Word's List.

2. Data Cleaning: It removes useless words from the Word's List; these useless words are stored in a stop words database as appear in the figure. The database has 311 stop words with a size 3KB.

3. Log Files Analysis: It removes irrelevant records from Log file. In order to enhance the efficiency of usage based retrieval algorithm by a useful records only. Log file Analysis consist a series of process like data cleaning, user identification, session identification.

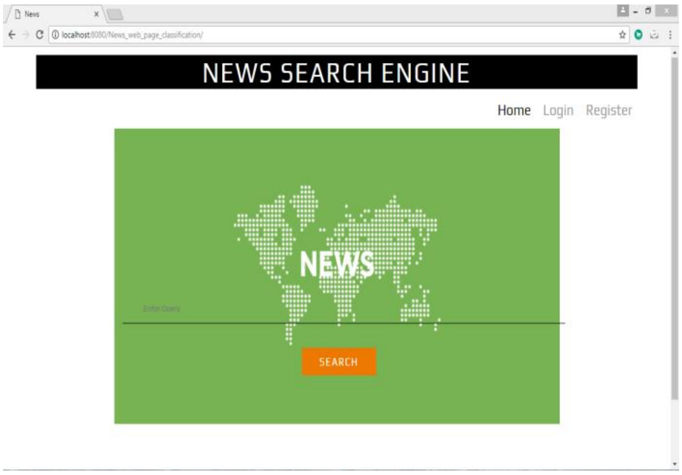
4. Indexing: Indexing is a process for describing or classifying a document by index terms; index terms are the keywords that have meaning of its own (i.e. which usually has the semantics of the noun). This index terms are grouped in an indexer and stemmer is service this stage by improving the group of these keywords in the indexer.

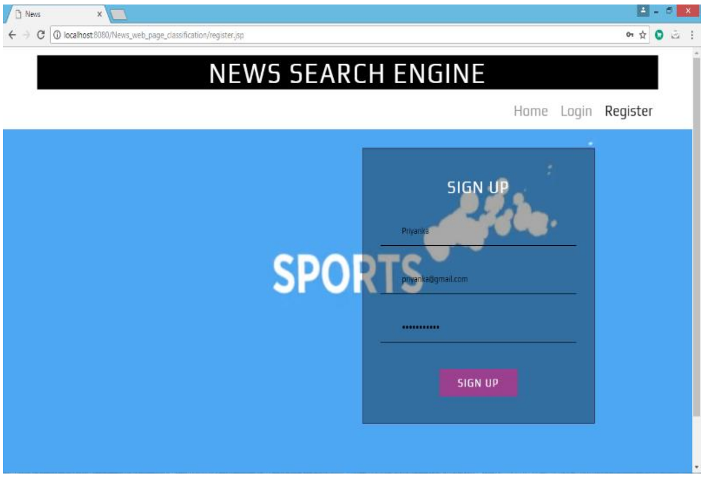
5. Content-Based Ranking: The user's query is matched with the index terms to get the relevant documents to the query. Documents are then ranked using ranking algorithms according to the most relevant to the user's query.

6. Usage-Based Parameters: In this stage we calculate several parameters which are the inputs to our algorithm.

7. Usage-Based re-ranking: It's the combination of the pervious modules to provide a new weight called usage based weight for the pages, then ranking those pages according to their new weight.

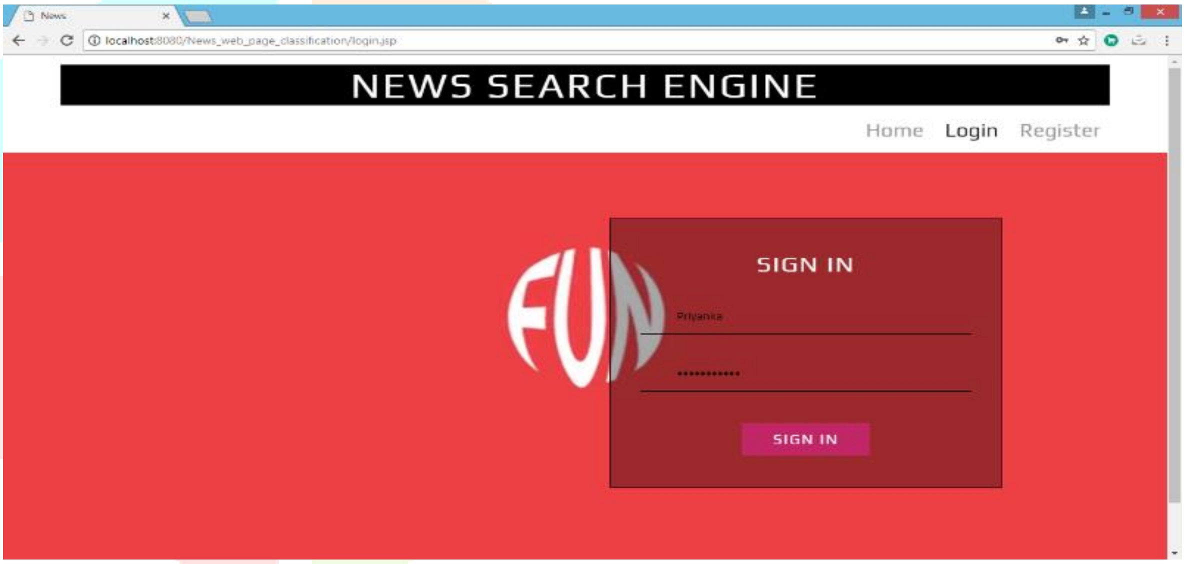
VII. RESULTS





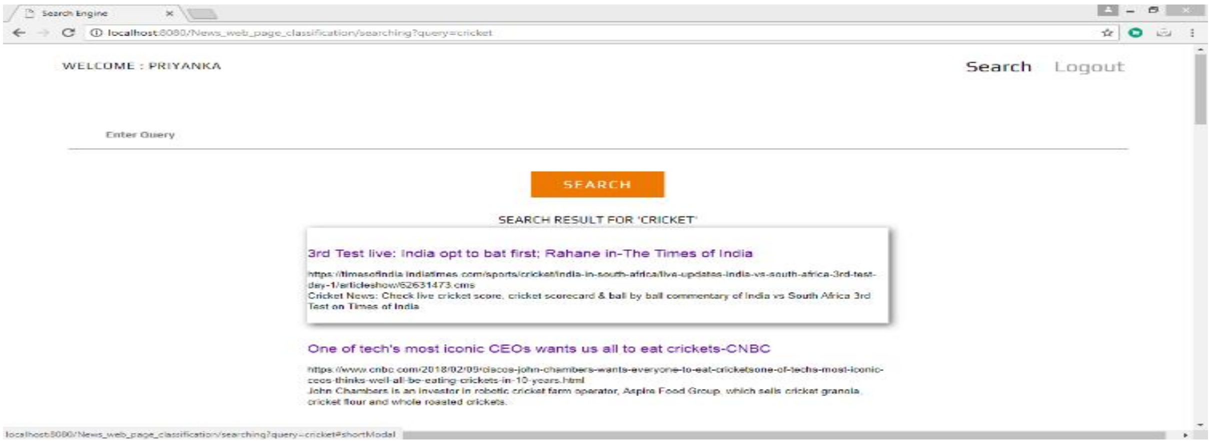
Priyanka
priyanka@gmail.com
....

Figure 2. Homepage **Figure 3. Registration Page**



Priyanka
....

Figure 4. Login Page


Search Engine
localhost:8080/News_web_page_classification/searching?query=cricket
Search
Logout
Enter Query
SEARCH
localhost:8080/News_web_page_classification/searching?query=cricket#shortModal

WELCOME : PRIYANKA

SEARCH RESULT FOR 'CRICKET'

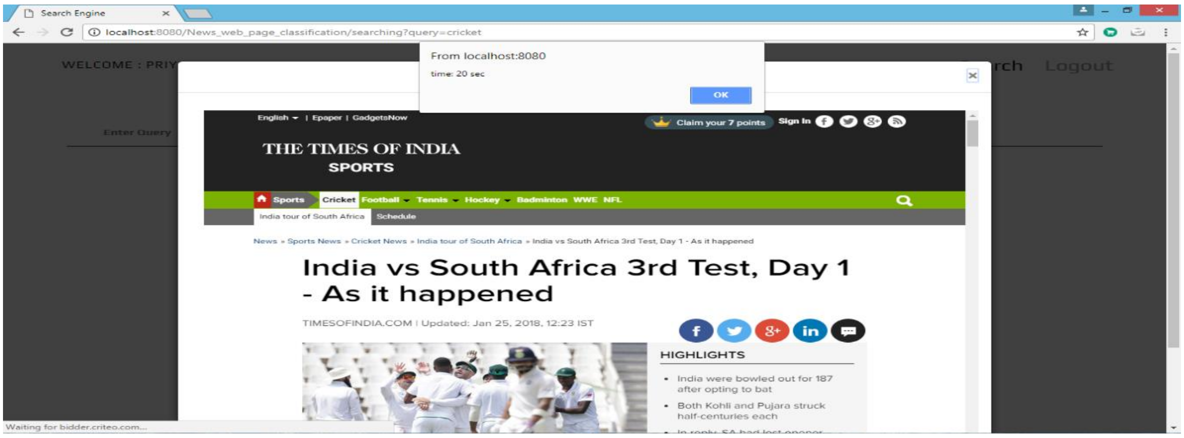
3rd Test live: India opt to bat first; Rahane in-The Times of India

https://timesofindia.indiatimes.com/sports/cricket/india-in-south-africa/live-updates-india-vs-south-africa-3rd-test-day-1/articleshow/62631473.cms
Cricket News: Check live cricket score, cricket scorecard & ball by ball commentary of India vs South Africa 3rd Test on Times of India

One of tech's most iconic CEOs wants us all to eat crickets-CNBC

https://www.cnbc.com/2018/02/05/ciscos-john-chambers-wants-everyone-to-eat-cricketsone-of-techs-most-iconic-ceos-thinks-well-all-be-eating-crickets-in-10-years.html
John Chambers is an investor in robotic cricket farm operator, Aspire Food Group, which sells cricket granola, cricket flour and whole roasted crickets.

Figure 5. Search Query Result


Search Engine
localhost:8080/News_web_page_classification/searching?query=cricket
WELCOME : PRIY
rch
Logout
Enter Query
English | Epaper | GadgetsNow
Claim your 7 points
Sign In
THE TIMES OF INDIA
SPORTS
Sports
Cricket
Football
Tennis
Hockey
Badminton
WWE
NFL
India tour of South Africa
Schedule
News » Sports News » Cricket News » India tour of South Africa » India vs South Africa 3rd Test, Day 1 - As it happened
HIGHLIGHTS
OK
Waiting for bidder.criteo.com...

From localhost:8080

time: 20 sec

India vs South Africa 3rd Test, Day 1 - As it happened

TIMESOFINDIA.COM | Updated: Jan 25, 2018, 12:23 IST

- India were bowled out for 187 after opting to bat
- Both Kohli and Pujara struck half-centuries each

Figure 6. News Article

VIII. Conclusion

The emergent field of online newspaper exhibits a rich region which can advantage significantly from automatic classification approach. This paper exhibited a compelling way to deal with understand the automatic news web page classification based on content attributes, structure attributes and URL attributes of news web pages. We begin with the perception that the best possible decision of attributes can have a significantly affect the performance of classification algorithm. We extract the attributes from ten different websites. We used Naïve Bayes algorithm for classification and conducted comparative experiments with various existing algorithms on the same dataset from ten different websites, and the results show that Naïve Bayes perform better than other algorithms; this algorithm provides adequate classification accurateness with the different news datasets. There are several possible extensions to this study. Our future target is to explore a technique to identify and remove the irrelevant information of comments section. We also consider the other structural attributes of news web pages for better classification accuracy.

References

[1] X. Wu, G.Q. Wu, F. Xie, Z. Zhu, and XG. Hu, "News filtering and summarization on the web," IEEE Intelligent Systems 5, pp.68-76, 2010.

[2] K. Sarkar, M. Nasipuri, and S. Ghose, "Machine learning based keyphrase extraction: comparing decision trees, Naive Bayes, and artificial neural networks," Journal of Information Processing Systems 8.4, pp. 693-712, 2012.

[3] D.D.C. Reis, P.B. Golgher, A. S. Silva, "Automatic web news extraction using tree edit distance," Proceedings of the 13th international conference on World Wide Web. ACM, 2004.

[4] S. Tongchim, V. Sornlertlamvanich, and H. Isahara. "Classification of news web documents based on structural features,"Advances in Natural Language Processing. Springer Berlin Heidelberg, pp. 153-160, 2006.

[5] L. K. Shih, and D. R. Karger. "Using urls and table layout for web classification tasks," Proceedings of the 13th international conference on World Wide Web. ACM, 2004.

[6] M.Y. Kan, and H.O.N. Thi. "Fast webpage classification using URL features," Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005.

[7] A. Selamat, and S. Omatu. "Web page feature selection and classification using neural networks," Information Sciences, 158, pp. 69- 88, 2004.

[8] A. Hotho, A. Maedche, and S. Staab. "Ontology-based text document clustering," KI 16.4, pp. 48-54, 2002.

[9] A. K. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y .Ng, "Improving text classification by shrinkage in a hierarchy of classes," In J. W. Shavlik, editor, Proceedings of the 15th Intl. Conference on Machine Learning, ,Madison, US, Morgan Kaufmann Publishers, San Francisco, US, pp. 359-367, 1998 .

[10] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," In Proceedings of the 14th Intl. Conference on Machine Learning, pp. 170–178, 1997.

[11] R. Agrawal and R. Srikant, "On integrating catalogs," In Proceedings of 10th Intl. Conference on the World Wide Web, Hong Kong, CN, ACM Press, New York, US, pp. 603-612, 2001.

[12] D. Billsus and M. J. Pazzani, "A hybrid user model for news story classification," In Proceedings of the Seventh Intl. Conference on User Modeling, Springer-V erlag New Y ork, Inc., pp. 99-108, 1999.

[13] D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," Machine Learning: ECML-98, pp. 4–15, 1998.