Based on the user query, search engine results are retrieved. Every result is individually analyzed based on keywords and content. The particular score is awarded to every result on the basis of keyword present in the title, count of keyword in web page and total number of words present in the web page. The total relevancy of the particular link against user request is computed by summarizing all the scores of the keyword and content words. Finally, all these results are re-ordered by arranging them with their decreasing order of scores. Thus the result now present on the top is the most relevant according to the user query.

4.2 Usage Mining

User clicks any one of the links presenting to him/her after re-ordering. The query user enters, url, no. of clicks and the contents retrieved from that web page are stored on the server log, which represents user's interest for that particular query. When next time user will enter any other or same query as entered before, the scores will be awarded on the basis of data stored in server log. And again the re-ordered list of results will be presented to the user. Thus the result then present on the top is the most relevant according to user's interest.

5. Specification

Implementation is the stage of project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and its constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.
The proposed system is divided into four parts.

5.1. Data Extraction

Data extraction is the first stage in this project. Since we are concerned only about the content on the websites, we use web content mining to extract the data. Web content mining is very similar to text mining and data mining. In web content mining the data is either semi-structured or unstructured whereas in data mining the data is more structured. Due to the exponential growth in web content and its usage, many applications are built to use this data and present the data in a user-friendly manner. There are various problems involved in extracting data from the Internet due to lack of a standard structure used by websites to store the data. When user enters the query in the system API, the query is forwarded to the search engine and the top n results are extracted from the search engine.

5.2. Preprocessing

Pre-Processing is an important step in text based mining. Real world data tend to be dirty, incomplete and inconsistent. Data pre-processing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process.

5.3. Page Score

(a) by content mining
For re-ordering the results, every result link is given the specific page score. The page score is depend on the number of times the keyword in the user query appears in the web content and the number of total contents on the web page. Some additional weight is given if the keyword in the user query appears in the title. The value of score of each result is sorted in descending order to get the most relevant content for the user query. Re-ordered results are sent back to the user so that the top most page is more relevant for the user query. When user clicks any of the result, the url, query and contents extracted from that web page are stored in the web page. The contents are extracted by NLP using POS tagger.

- **Natural Language Processing**
  NLP provides means of analyzing the text. The goal of NLP is to make computers analyze and understand the languages that humans use naturally. As the reader has probably already deduced, the complexity associated with natural language is especially key when retrieving textual information to satisfy a user's information needs. This is why in Textual Information Retrieval, NLP techniques are often used both for facilitating descriptions of document content and for presenting the user's query, all with the aim of comparing both descriptions and presenting the user the documents that best satisfy their information needs

- **Part of Speech Tagging**
  When the user clicks the result link he is interested in, the web page gets open in the browser. The contents present on the web page are tokenized and tagged using POS tagging. A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns part of speech to each word (and other token), such as noun, verb, adjective, etc., Following is the list of POS. Out of all these only noun and adjectives are stored into server log.

b) by usage mining

When user enters the query it is first check for the server log entries. If the log contains some entries for previously accessed data then the page score is given on the basis of data present in the server log. Presence of same url and more number of clicks for any url gives additional score to that particular result.

5.4. Server Log Maintenance

For storing the history of user's access behaviour, the server log is maintained. The log contains query entered by user, the URL of the selected result, the contents retrieved from the selected page and number of times he made the click on that particular URL. The log is updated automatically whenever the user selects the web page. The WAMP server has been used to maintain the server log.

6. Experimental Results

Experiment is conducted with a user query "sun". Top 10 results are retrieved from the search engine. Following table represents results retrieved from search engine.

Table 1: Top 10 results retrieved for query "sun

S.No	ID	URL
1	SR1	http://www.thesun.co.uk/
2	SR2	http://en.wikipedia.org/wiki/su
3	SR3	http://www.oracle.com/us/sun
4	SR4	http://www.torontosun.com/
5	SR5	http://nineplanets.org/sol.html
6	SR6	http://www.calgarysun.com/
7	SR7	http://www.vancouvarsun.com/
8	SR8	http://www.baltimoresun.com/
9	SR9	http://edmotionsun.com/
10	SR10	http://www.sun-sentinel.com/

Keyword and content based ranking approach is applied and results are reordered. At the present stage there is no data in the server log.

These re-ordered result list was presented to the user and asked to select any one of the results. The user was totally unaware about the proposed and objectives of the project. User entered his choice. The required data is now stored in the server log. Depend on user's choice the results are now re-ordered by combine approach of usage and content mining. The order of the results provides by search engine and proposed system is compared by using manual ranking from user. Table 2 shows this comparison.

ID	*Search Engine Ranking*	*Content Based Ranking*	*Content and Usage Based Ranking*	*Manual Ranking*
SR1	1	9	6	6
SR2	2	1	2	2
SR3	3	10	10	10
SR4	4	7	7	7
SR5	5	2	3	3
SR6	6	8	9	8
SR7	7	3	4	4
SR8	8	4	1	1
SR9	9	6	8	9
SR10	10	5	5	5

6.1 Performance Evaluation

Performance evaluation of the proposed approach and the search engine is done based on Precision measure.
Precision measure is calculated based on the following formula.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Where,
tp – True Positive (Correct result)
fp – False Positive (Unexpected Result)

Table 2 represents the matching of manual ranking against proposed approach ranking and search engine ranking. Document SR1, 2, 5, 7, 8, 9 represents the mismatching of manual ranking against content based ranking and documents SR6, 9 represents the mismatching of manual ranking against content and usage based ranking.

Table 3: Precision Measure

Different Methods	tp	fp	Precision
Search Engine Ranking	2	8	0.2
Content Based Ranking	4	6	0.4
Content & Usage Based Ranking	8	2	0.8

From the table 3, it is understood that precision of the search-engine is 0.2, precision of content mining is 0.4 and precision of combine approach of content and usage mining is 0.8 out of 1. The results of the performance measure are plotted in Figure 3.

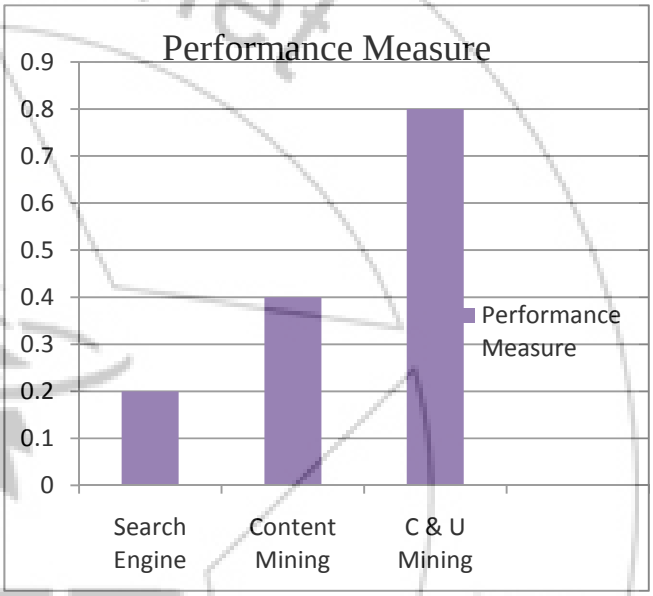

Figure 3: Performance Measure

7. Conclusion

Among the various techniques proposed for organizing and re-ordering of search results from the web, as described in the literature review, combination of content and usage mining to re-ordering the search results was selected for the implementation of proposed work to provide useful information to the user.

To gather the useful information from the web page, the user interested in, the log is maintain on the database of WAMP server. From this server log, interest of particular user can be found easily.The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience which can be achieved by web usage mining while web content mining removes persistence problem. This work is the combination of content and usage mining which works better than using any one of them. The experimental results demonstrate that proposed model is promising. The proposed approach provides a solution to minimize the persistence and incorrect information problem in a single model. The findings can be applied to the design of Crawling Systems and for Discrimination Prevention. The

system also has contribution to the field of Information Retrieval.

7.1 Future Work

The proposed work is focused only on re-ordering the search results, retrieved from search engine, according to the user's interest and can be further extended to work for searching the results according to the user interest. Proposed work is based on text based mining. As the information on web is present in the images, audio and video formats also, the future work can be focused on all these different formats.

References

[1] R. Cooley, B. Mobasher and J. Shrivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, pp. (ICTAI), 1997.
[2] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.
[3] I. Mele, " Web Usage Mining for Enhancing Search – Result Delivery and Helping Users to Find Interesting Web Content," ACM SIGIR Conf. Research and Development in Information Retrival (SIGIR '13), pp. 765-769, 2013.
[4] P. Sudhakar, G. Poonkuzhali, R. Kishor Kumar, "Content Based Ranking for Search Engines", Proc. International Multi Conference of Engineers and Computer Scientists (IMECS 12), 2012.
[5] Z. Lu, H. Zha, X. Yang, W. Lin, Z. Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions," Proc. IEEE Transactions on Knowledge and Data Engineering, pp. 502-513, 2013
[6] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining -Concepts, Applications & Research Directions", University of Minnesota, USA, Chapter 3. Pg 52-71.
[7] Gibson, J., Wellner, B., Lubar, S, "Adaptive web-page content identification", In WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management. New York, USA,2007.
[8] Georgies Lappas, An overview of web mining in societal benefit areas, The 9th IEEE International Conference on E-Commerce Technology, IEEE 2007.
[9] U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 391-400, 2005.
[10] O. Zamir and O. Etzioni, " Web Document Clustering: A Feasibility demonstration," ACM (SIGIR, 99) , pp. 46-54.
[11] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.
[12] J. Zhu, J. Hong, and J. G. Hughes. Pagecluster: "Mining conceptual link hierarchies from web log files for adaptive web site navigation," ACM Trans. Internet Technol., 4(2), 2004.
[13] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-ow graph: model and applications. In CIKM, 2008.
[14] R. W. White, M. Bilenko, and S. Cucerzan. "Studying the use of popular destinations to enhance web search interaction,". In SIGIR, 2007.
[15] Y. Xie and D. O'Hallaron, "Locality in search engine queries and its implications for caching," In IEEE Infocom 2002, pages 1238{1247, 2002.
[16] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl., 1(2):12, 23, 2000.

Author Profile

Ms. Shital C. Patil received B.E. degree in Computer Science and Engineering from Sant Gadge Baba Amravati University in 2012. Presently in pursuing M.E second year Computer Science & Engineering.