



Exploring Structure Oriented Feature Tag Weighting Algorithm for Web Documents Identification

Karunendra Verma¹(✉), Prateek Srivastava¹, and Prasun Chakrabarti²

¹ Department of CSE, Sir Padampat Singhania University, Udaipur, India
k.verma2006@gmail.com, prateek.srivastava@spsu.ac.in
² Department of Computer Science and Engineering, ITM Universe Vadodara,
Paldi 391510, Gujarat, India
dean.research@itmuniverse.ac.in

Abstract. There are various ways of web page classification but they take higher time to compute with lesser accuracy. Hence, there is a need to invent an efficient algorithm in order to reduce time and increase web page classification result. It is generally find that a few tags like title can contain the principle substance of text, and these patterns may have an impact on the adequacy of text classification. Although, the most widely recognized text weighting calculations, called term frequency inverse documents frequency (TF-IDF) doesn't consider the structure of website pages. To take care of this issue, another feature tags weighting calculation is put in advanced. It thinks about the web page structure data like title, Meta tags, head etc. also content the useful information. In this proposed study first web site pages data are pre-processed and find text weight using TFIDF, after that using feature tag weighting calculation, frequent and important tags will find; then on the basis of text weight and tags weight, web document will classify.

Keywords: Web page classification · Feature tags · TFIDF · Text weight

1 Introduction

Presently day by day Internet has turned out to be extremely well known and intuitive for exchanging data. The web is tremendous, differing and dynamic thus increase the versatility, interactive media information and fleeting issues. The development of the web has its result in a gigantic measure of data that is currently unreservedly offered for client's entrance. A few various types of information must be taken care of and composed in a way that they can be gotten to by the clients viably and productively.

The web is an accumulation of interrelated documents on at least one web servers. Web mining is the utilization of information mining methods to extricate learning from web information including web archives, hyperlinks between pages, usages logs of sites and so on.

Web mining is comprehensively partitioned into following classes: web content mining, web usage mining and web structure mining.

Text classification is a procedure of partitioning text into one or multiple classes. As the advancement of the web technology, objective is to achieve accurate web text

from web documents. Web documents have clear identifier (i.e., HTML tags) to convey its structure information. Generally, in the content extricating process, HTML page structures are expelled and separate plain text from each website pages. In many cases, there is a lot helpful data regarding the content organization based on HTML tags. Numerous investigators demonstrate to the structural data, particularly HTML tags, similar to table design, hyperlink, be able to utilized to enhance viability of web content classification.

2 Review of Literature

Kovacevic et al. [9] proposed hierarchical representation that incorporates browser screen which facilitates with each HTML article in a page. Utilizing image data one can characterize heuristics for the acknowledgment of basic page zones, for example, footer, header, right and left menu, and focus of a web page. In the underlying examinations the creator demonstrates that utilizing heuristics characterized objects are perceived legitimately in 73% of cases. At long last, demonstrate that a Naive Bayes classifier, given that the proposed representation.

Zou et al. [16] have discovered that because of the proximity of the raucous information here is a requirement for characterization of the web page for true applications. A strategy which will appropriately guarantee the arrangement be the support vector machine since it has the ability of speculation. Creator's recommended strategy gives a way which will expand the precision of arrangement by joining the support vector machine idea among the K - nearest neighbor procedures.

Tomar et al. [4] present the idea of an order device for pages called Web Characterize, which utilizes changed customized naive Bayesian calculation with a multinomial form to arrange pages hooked on different classifications. In this exploration test result alongside the grouping exactness investigation with expanding vocabulary measure, was likewise appeared.

Ryan et al. [10] examined the region of classification arrangement has an accentuation on recovering the highlights, for example, content from the particular archives. Since the principle point of work is considered whether visual properties of HTML site page can altogether enhance the arrangement of pulverous sorts. Evidently, it appears that it would put a noteworthy test and will be likewise helpful to recover those visual attributes which getting the design highlight of types. The majority of site pages delivered from different business sites and physically sorted into types. The three unique qualities are thought about one next to the other (a). With the literary attributes (b). With the HTML qualities (c). Visual qualities. Creator's work can demonstrate that by utilizing HTML qualities and URL attributes helps in expanding the precision of characterization when contrasted with printed alone. In this way, it additionally appears that by including the visual attributes, it builds the pulverous grouping.

Kang et al. [7] exhibit an investigation on mining web information from the various accessible information on WWW. As the pages are not completely organized so it ends up noticeably hard deciding from the useful block techniques which give the valuable information extraction from the futile information, for example, promotions which is more vital. In this proposed strategy creator present a website page arrangement in type

of pieces by building a tree arrangement demonstrate that show the HTML include and a vector display that speaks to an element of blocks. Hence, by building the single classifier it ends up noticeably hard to characterize a piece precisely. To defeat this issue in proposed strategy creator utilizes the various classifiers one for each preparation informational collection and characterization technique prevails by consolidating every one of them.

Mun et al. [11] found that the size of web page increases a lot as the number of offered services as well as link increases and then due to their accessing speed decreases. The author uses the link graph arrangements for troubleshoot this problem. By introducing this link graph system author enables to reduce the load of server to a greater extent.

Rathod [13] indicates frameworks of three unique methods of web pages mining, in particular web structure mining, web usage mining and web content mining. The advancement and utilization of Web mining strategies with regards to web content usage and structure information will prompt substantial enhancements in numerous web applications as of web crawlers and web specialists to web examination and personalization.

Gowri et al. [3] portrayed a short overview about the current approach in web administrations synthesis. The principle looks into regions in web administrations are identified with revelation, security, and creation. Among every one of these regions, web administrations organization ends up being a testing one in light of the fact that inside the administration arranged figuring area, Web benefit synthesis is a successful acknowledgment to fulfill the hurriedly changing prerequisites of business. In this manner, the Internet benefit creation has unfurled itself extensively in the exploration side. Be that as it may, the present endeavors to order Web benefit structure are not fitting to the targets. This article proposes a novel categorization matrix for Web service work, which recognizes the unique situation and innovation measurements. The setting measurement is gone for examining the QoS effect on the exertion of Web benefits creation, although the innovation measurement concentrates on the system impact on the exertion. At last, this paper gives a proposal to enhance the nature of administration determination which takes part in the arrangement procedure with C skyline approach utilizing operators.

Sarac et al. [14] worked on the firefly algorithm (FA) inspired by the flashing behavior of fireflies, which belongs to the category of Meta heuristic algorithm. It flashes primary intention to attract other fireflies through a signaling system.

Jain et al. [2] proposed another strategy "Intelligent Search Method (ISM)". In this technique creator proposes to index the web pages via an intelligent search approach. This new strategy incorporated with any of the page positioning calculations to deliver better and significant indexed lists.

Keller et al. [8] introduce a GRABEX strategy for removing navigational block pieces in light of the connection designs. The technique was connected to mine breadcrumb routes. Dissecting to which additional navigational chunk type the GRABEX strategy can be connected is additionally intriguing for prospect work. A creator trusts that paginations or past/next routes can be mined too if appropriate graph creation strategies are actualized. The GRABEX strategy can likewise be reached

out to extract non-navigational page components if diagrams are not produced from hyperlinks but rather from different structures e.g. text or linked images.

Jose et al. [6] demonstrate the Rough set hypothesis applications in different areas like company, prescription, trade, media transmission and numerous different fields. The consequences of this approach can be utilized for target promoting on the grounds that sponsors can post their notices on content pages particularly pages in bring down estimate. This likewise distinguishes the most favored substance by a client since clients invested more energy in potential pages.

Ye et al. [15] enhanced and proposed a kind new technique of semantic relevancy algorithm based for semantic importance calculation in light of the Wikipedia hyperlink arrange, incorporating the semantic data in the paging system and the class organize sensibly to complete semantic relationship figuring.

Sadegh et al. [1] explored social tags as a novel confirmation to categorize objects on the web. A new linkage structure between objects and tags is investigated for categorization. Tags moreover work as bridges to attach the heterogeneous domains of objects.

He et al. [5] work in view of the way that the web is an accumulation of different web records. The grouping of a web record is implied for three things for the most part: indexing, search and retrieval. There is a distinction between web grouping and content characterization. This distinction is because of the structure of the web reports. These distinctions could be at least one of the accompanying: meta information, the title of the record and different connections accessible in the archive and so on. In this paper, creators have picked both of the accompanying strategies, for example, Information gain and χ^2 - test for feature selection for classification. After feature selection, this paper uses Support Vector Machine (SVM) classifier for categorization. The strategy affirms, evaluates and broadens past study by presenting another structure-based technique for depiction and order of web archives. Contrasted with conventional web archive order strategies, consolidating the complete text among structure Information gain almost 6% exactness change on account of comparative classifications and 3.7% correctness enhancement in the case of different categories.

Qian et al. [12] worked on novel weighted Hamming distance based on Page Rank algorithm for anomaly intrusion detection. Using the Hamming distance with the Page Rank weight to estimate deformity degree of unusual system calls and focus on optimizing the algorithm complexity.

Chen et al. [19] characterized five best level type classifications and grew new techniques to mine 31 features from web database, which examined both features and contents. Their assessment results comes that extra features can help a classifier enhances its knowledge of the categorization.

Abramson et al. [20] exhibited a technique to facilitate uses data from URLs for website page database since a few URLs may include some text to shows the class. This approach can partially take care of the issue; however it is as yet not a general approach for all web pages genre.

3 Research Gap

The literature review entails that; the classification done on the dataset of web structure is optimized by structure based web document analysis. However, beside these described techniques there are various other ways also to perform web structure based classification. Certainly, web structure based classification gives better result in association with feature selection results because it finds various features in the record of dataset. But while using this technique with simple web structure based classification, there is a scope of improvement in the following two concerns.

- Web structure based classification itself takes longer time to compute.
- The result of simple web structure based classification is not that much optimized.

And the reasons behind these two concerns are the accuracy of the classification method. It is proposed to improvise these concerns to get better efficiency in this work.

4 Methodology and Used Algorithms

To overcome above limitations, work has divided into following steps (Fig. 1):

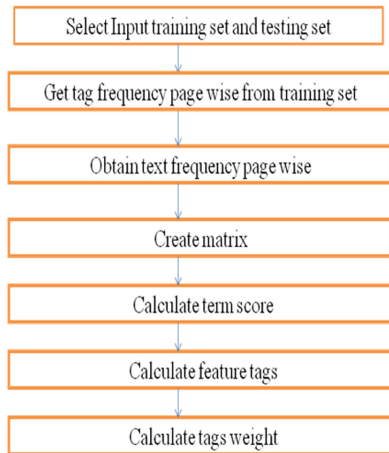


Fig. 1. Work flow

4.1 Term Frequency Inverse Document Frequency (TFIDF) [21]

$$TermScore_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \times \log \frac{|D|}{|\{d : t_i \in d\}|} \quad (1)$$

Where

- n_{ij} , no. of presences of term t_i in page d_j ;
- n_{kj} , sum of presences of all term in page d_j ;
- D , total number of pages;
- d , number of pages which incorporated term t_i .

4.2 Feature Tag Weighting Algorithm

Tag Frequency [18]

$$tf_i(t, d) = \sum_{i \in P} \left[tf_t(t_i, d) \times \frac{a_i}{\sqrt{\sum_{j=0}^k a_i^2}} \right] \quad (2)$$

Where

- $tf_t(t, d)$, tag frequency of term t in page d ;
- $tf_t(t_i, d)$, tag frequency of the term t in tag i ;
- a_i is the tag weighting coefficient and $i \in P$ and P is the set of tags.

Tag weight [18]

$$W_t(t, d) = \frac{tf_t(t, d) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in d} [tf(t, d) \times \log(N/n_t + 0.01)]^2}} \quad (3)$$

Where

- $W_t(t, d)$ is feature tag weighting of term t in page d ;
- $tf_t(t, d)$ is frequency of the word t in page d ;
- N is total number of pages;
- n_t is number of pages which included term t .

5 Programming Environment and Results

To simulate above work we used Net Beans IDE 8.2 and JDK 1.8. Investigations utilize the Bank Search dataset [17], which is particularly intended to help an extensive variety of web pages processing tests. The database comprises of 2202 web archives arranged into ten uniformly sized classes like A: Commercial banks Banking and finance, B: Building society Banking and finance, C: Insurance agencies Banking and finance, D: Java Programming languages, E: C++/C Programming languages, F: VB Programming languages, G: Astronomy Science, H: Biology Science, I: Soccer Sport, J: Motor etc. and each contains 200 web archives (Figs. 2, 3, 4, 5, 6, 7 and 8).

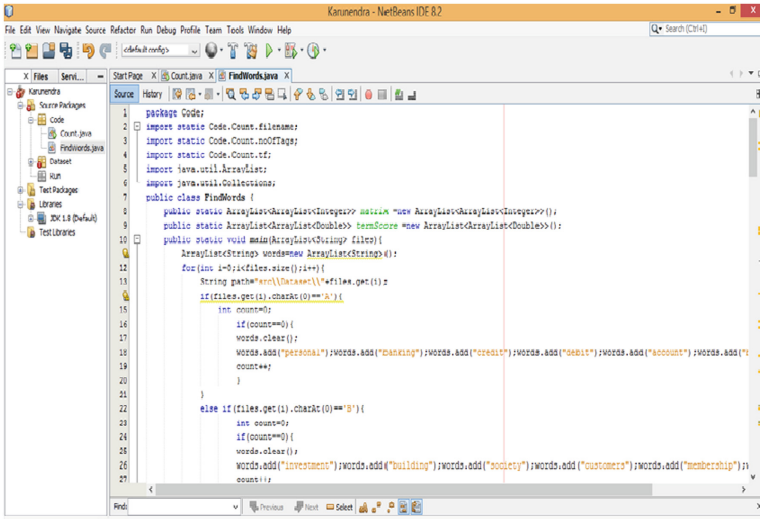


Fig. 2. Programming environment

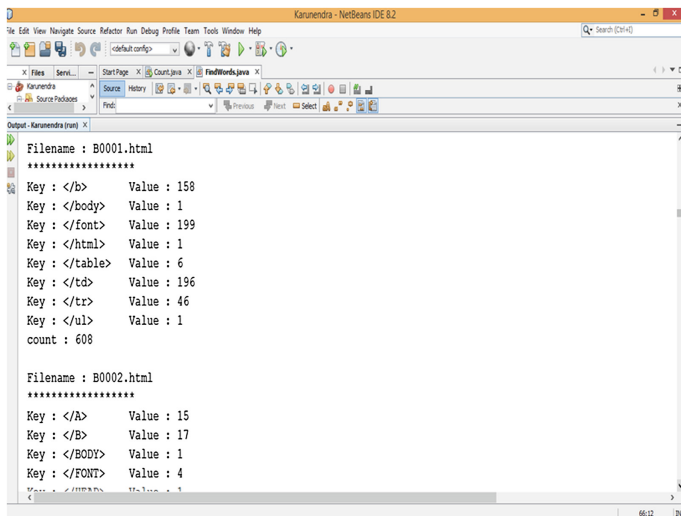


Fig. 3. Tag frequency calculation

The results include characterization of two classes from the dataset. The initial 1500 records are utilized as training set and the rest 500 records are utilized as testing set. a few classes are very similar, while a few classes are very distinct (e.g. class A: Commercial Banks and J: Motor Sport). Categorizing related classes is obviously a new troublesome machine-learning task.

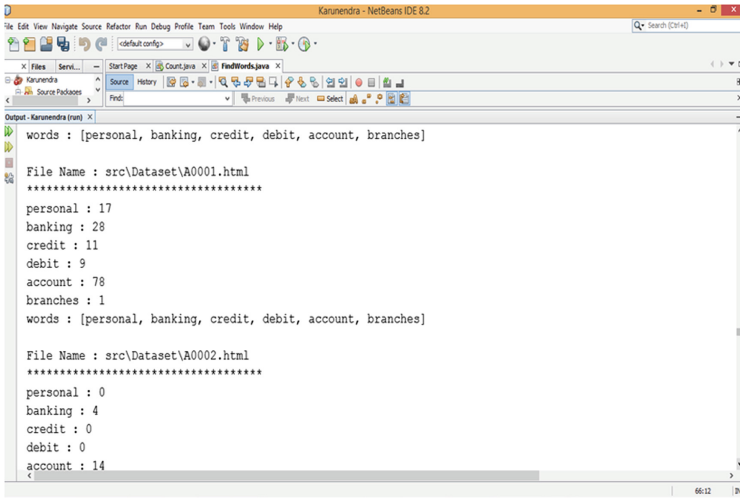


Fig. 4. Text frequency calculation

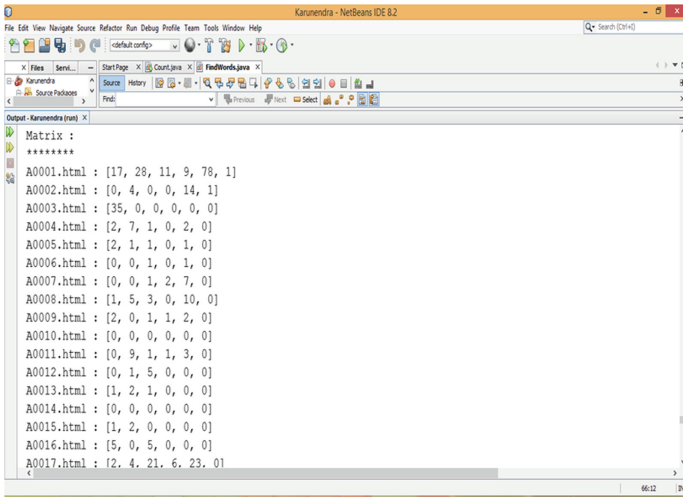
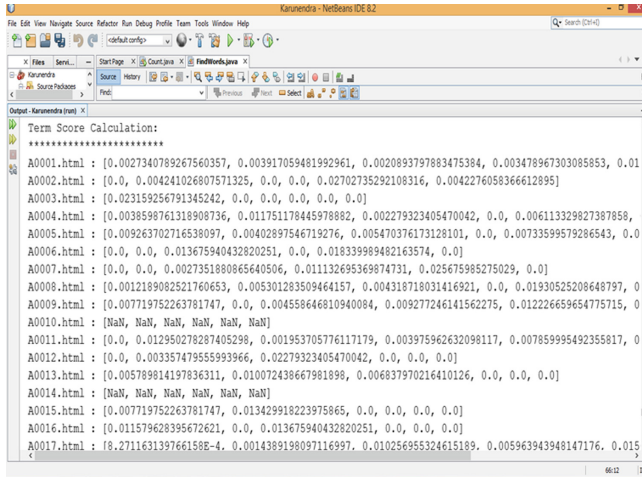


Fig. 5. Matrix creation

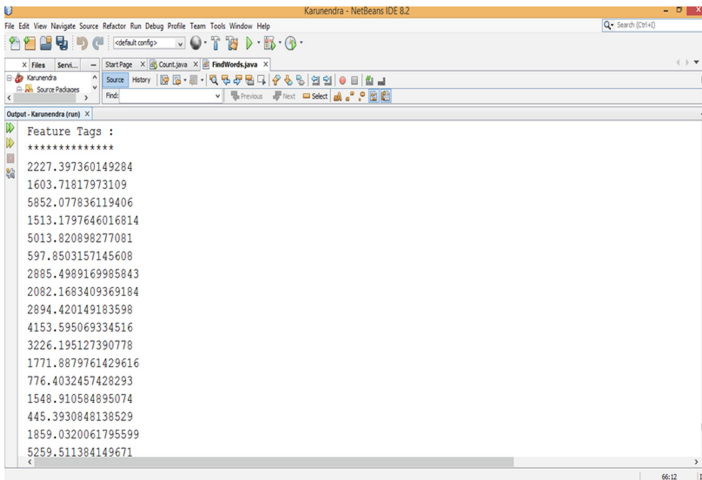


```

Karumendra - NetBeans IDE 8.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
StartPage Count.java FindWords.java
Karumendra
Source History
Output - Karumendra (run)
Term Score Calculation:
*****
A0001.html : [0.0027340789267560357, 0.003917059481992961, 0.0020893797883475384, 0.003478967303085853, 0.01
A0002.html : [0.0, 0.004241026807571325, 0.0, 0.0, 0.02702735292108316, 0.0042276058366612895]
A0003.html : [0.023159256791345242, 0.0, 0.0, 0.0, 0.0, 0.0]
A0004.html : [0.0038598761318908736, 0.011751178445978882, 0.002279323405470042, 0.0, 0.006113329827387859,
A0005.html : [0.009263702716538097, 0.00402897546719276, 0.005470376173128101, 0.0, 0.00733599579286543, 0.0
A0006.html : [0.0, 0.0, 0.013675940432820251, 0.0, 0.01833998482163574, 0.0]
A0007.html : [0.0, 0.0, 0.0027351880865640506, 0.01132695369874731, 0.025675985275029, 0.0]
A0008.html : [0.0012189082521760653, 0.005301283509464157, 0.004318718031416921, 0.0, 0.01930525208648797, 0
A0009.html : [0.007719752263781747, 0.0, 0.004558646810940084, 0.009277246141562275, 0.012226659654775715, 0
A0010.html : [NaN, NaN, NaN, NaN, NaN, NaN]
A0011.html : [0.0, 0.012950278287405298, 0.001953705776117179, 0.003975962632098117, 0.007859995492355817, 0
A0012.html : [0.0, 0.00335747955593966, 0.02279323405470042, 0.0, 0.0, 0.0]
A0013.html : [0.005789814197836311, 0.010072438667981898, 0.006837970216410126, 0.0, 0.0, 0.0]
A0014.html : [NaN, NaN, NaN, NaN, NaN, NaN]
A0015.html : [0.007719752263781747, 0.013429918223975865, 0.0, 0.0, 0.0, 0.0]
A0016.html : [0.011579628395672621, 0.0, 0.013675940432820251, 0.0, 0.0, 0.0]
A0017.html : [8.271163139766158E-4, 0.0014389198097116997, 0.010256955324615189, 0.005963943948147176, 0.015
66/2 7/6

```

Fig. 6. Term score calculation



```

Karumendra - NetBeans IDE 8.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
StartPage Count.java FindWords.java
Karumendra
Source History
Output - Karumendra (run)
Feature Tags :
*****
2227.397360149284
1603.71817973109
5852.077836119406
1513.1797646016814
5013.820898277081
597.8503157145608
2885.4989169985843
2082.1683409369184
2894.420149183598
4153.595069334516
3226.195127390778
1771.8879761429616
776.4032457428293
1548.910584895074
445.3930848138529
1859.0320061795599
5259.511384149671
66/2 7/6

```

Fig. 7. Feature tags calculation

In “Fig. 9,” four major tag types in a web pages (title, meta, link, and image) importance were compared from rest of all tags. Then this number was normalized against the web pages with respect to tag weighting function.

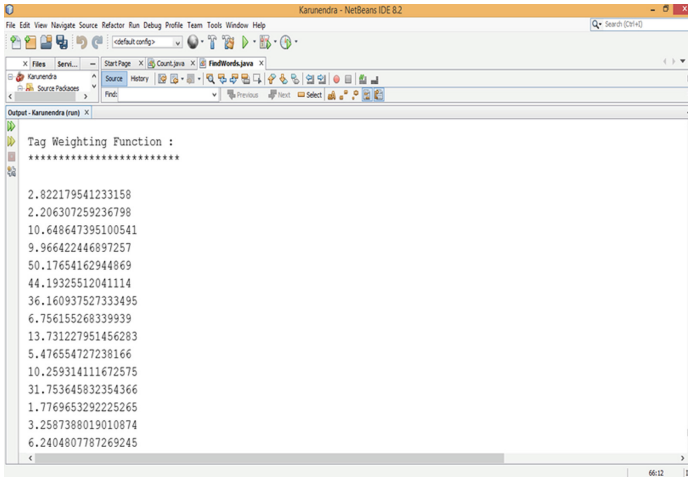


Fig. 8. Tag weight calculation

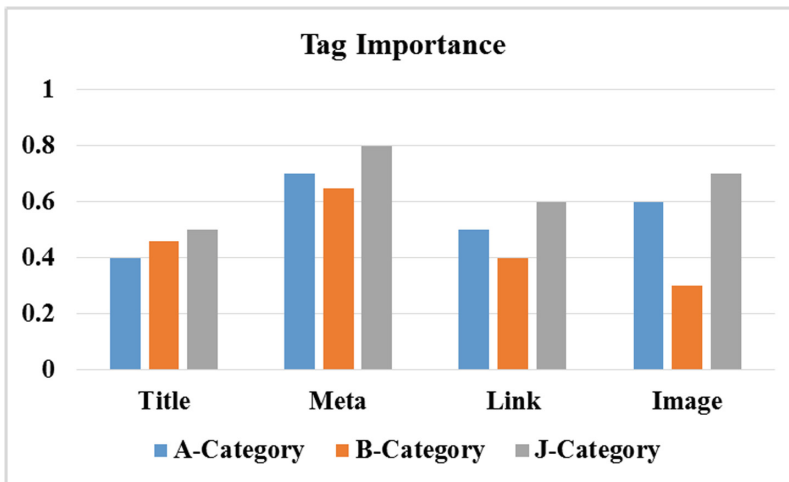


Fig. 9. Tag importance

6 Conclusion and Future Work

This article has exhibited a structure-based strategy for fabricating high precise web page categorization. It has shown in “Fig. 9”, the handiness of thinking about structure data, which incorporates Links, META tags, TITLE and alternative texts of images. The process is assessed utilizing the Bank Search database, and the investigations show the reward of structure-based characterization for both similar categories and different categories. Pure text classification has not considered the difference in the web content, which has HTML tags to express further structure data of the content. In the event that

we simply utilize TF-IDF which suits to the text classification, web text might be overlooked. The feature tag weighting calculation considers the impact of HTML labels on the web content classification. It performs superior to TF-IDF in the impact of tagged web content classification. As indicated by our test result, include tag weighting calculation gets the good precisions values.

In the meantime, our test advises us that while characterizing the web text, we can also consider HTML tags with a specific end goal to enhance the impact of web content classification. Be that as it may, there are as yet some limitations in this paper like only some HTML tags are considered for results we may include some more for more accuracy of the results. Furthermore we can also apply the appropriate classifier for web pages categorization.

Acknowledgements. I would like to thank all the people those who helped me to give the knowledge about these research papers. I am thankful to Dr. Prateek Srivastava & Dr. Prasun Chakrabarti to encourage and guided in this topic which helped me to speed up the work for structure based web page classification for fast search. Finally, I like to acknowledge all the websites and IEEE papers which I have gone through and referred to create this research paper.

References

1. Sadegh, A.H., Hossein, R., Behroo, N.: Web page classification using social tag. In: IEEE International Conference on Computational Science and Engineering, vol. 4, no. 1, pp. 588–593 (2009)
2. Jain, A., Sharma, R., Dixit, G., Tomar, V.: Page ranking algorithms in web mining, limitations of existing methods and a new method for indexing web pages. In: International Conference on Communication Systems and Network Technologies, vol. 3, no. 1, pp. 640–645. IEEE (2013)
3. Gowri, R., Lavanya, R.: A novel classification of web service composition and optimization approach using skyline algorithm integrated with agents. In: IEEE Computational Intelligence and Computing Research (ICCIC), pp. 26–28 (2013)
4. Tomar, G.S., Verma, S., Jha, A.: Web page classification using modified naïve bayesian approach. In: IEEE TENCON 2006, Hong Kong, pp. 14–17 (2006)
5. Kejing, H., Henyang, C.: Structure-based classification of web documents using support vector machine. In: Proceedings of CCIS 2016, pp. 215–219. IEEE (2016)
6. Jose, J., Lal, P.S.: A rough set approach to identify content and navigational pages at a website, pp. 5–9. IEEE (2008)
7. Kang, J., Choi, J.: Block classification of a web page by using a combination of multiple classifiers. In: IEEE Networked Computing and Advanced Information Management, vol. 2, no. 1, pp. 290–295 (2008)
8. Keller, M., Hartenstein, H.: GRABEX: a graph-based method for web site block classification and its application on mining breadcrumb trails. In: WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), pp. 290–297. IEEE (2013)
9. Kovacevic, M., Diligenti, M., Gori, M., Milutinovic, V.: Recognition of common areas in a web page using visual information: a possible application in a page classification. In: IEEE Data Mining, pp. 250–257 (2002)

10. Ryan, L., Michal, C., Lei, Y.: Using visual features for fine-grained genre classification of web pages. In: Proceedings of the 41st Annual IEEE Hawaii International Conference on System Sciences, vol. 1, no. 10, pp. 7–10 (2008)
11. Mun, Y., Lee, M., Cho, D.: Classification of web link information and implementation of dynamic web page using Link Map System. In: IEEE Granular Computing, pp. 26–28 (2008)
12. Qian, Q., Li, J., Cai, J., Zhang, R., Xin, M.: An anomaly intrusion detection method based on PageRank algorithm. In: International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, pp. 2226–2230. IEEE (2013)
13. Dushyant, R.: A review on web mining. *Int. J. Eng. Res. Technol. (IJERT)* (2012)
14. Sarac, E., Ozel, S.A.: Web page classification using firefly optimization. In: IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) (2013)
15. Ye, F., Zhang, F., Luo, X., Xu, L.: Research on measuring semantic correlation based on the Wikipedia hyperlink network, pp. 309–314. IEEE (2013)
16. Zou, J.Q., Chen, G.L., Guo, W.Z.: Chinese web page classification using no se-tolerant up port vector machines. In: Natural Language Processing and Knowledge Engineering, IEEE NLP-KE, pp. 785–790 (2005)
17. Sinka, M.P., Corne, D.W.: BankSearch dataset (2005). <http://www.pedal.reading.ac.uk/bansearchdataset/>
18. Lu, Y., Peng, Y.: Feature weighting improvement of web text categorization based on particle swarm optimization algorithm. *J. Comput.* **10**(1), 260–269 (2006)
19. Chen, G., Choi, B.: Web page genre classification. In: Proceedings of the ACM Symposium on Applied Computing, pp. 2353–2357 (2008)
20. Abramson, M., Aha, D.M.: What’s in a URL? Genre classification from URL. In: Workshops at the 26th Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence, pp. 1–8 (2012)
21. Zhu, J., Xie, Q., Yu, S.I., Wong, W.H.: Exploiting link structure for web page genre identification. *Data Min. Knowl. Discov.* 1–26 (2015)