A New Weighted Keyword Based Similarity Measure for Clustering Webpages

Shihab Rahman, Dolon Chapa, and Shaily Kabir*
Department of Computer Science and Engineering
University of Dhaka
Dhaka, Bangladesh
*Email: shailykabir2000 {at} yahoo.com

Abstract— **Relevant information from the web can quickly be retrieved if logically similar webpages are grouped together. Indeed, the clustering of web pages makes entire group available to the user, thereby increasing the efficiency of web browsing. Nevertheless, clustering largely depends on the accuracy of similarity computation among the pages. In this paper, we propose a new weighted keyword based similarity measure for discovering the alikeness among the pages. We present each page using a vector of extracted keywords, which is then converted into a weighted vector by considering both frequency and position of the keywords in the page. For determining semantic similarity between two pages, we take into account both syntactic and semantic relatedness between the respective weighted vectors. Finally, the webpages are grouped using this similarity measure by applying a fuzzy clustering algorithm. Our experimental results based on different cluster validation indices show considerable improvement in page clustering as compared to use of other existing similarity measures.**

Keywords - web content mining; keyword extraction; similarity measure; fuzzy clustering; cluster validation index

I. INTRODUCTION

The World Wide Web has grown in a phenomenal rate in recent years. Moreover, web content has been changing every day with new information resulting in an enormous volume of semi-structured and unstructured data. This huge volume of data is quite useless to a user if he gets overwhelmed with hundreds of websites and webpages while searching and faces difficulties in extracting valuable information. In this context, web mining becomes relevant now-a-days to make the web more user friendly. Among three categories of the web mining, the content mining works with the unstructured and semi-structured data. Further, it aims to mine and extract useful information from the webpage content, and the extracted information is useful for grouping the similar webpages. In this context, our target is to group semantically related pages together for betterment of web accessing.

In this paper, we extend the notion of similarity between the webpages by considering their semantic as well as syntactic relatedness. We propose a new semantic similarity measure for computing alikeness among the pages. We present each page through a set of keywords, which are later transformed to a vector of weighted keywords while considering the frequency of the keywords along with their position in different tags within the page. We base our similarity computation between two pages by measuring relatedness between the respective weighted vectors of the keywords. A set of semantically related page models is generated by applying the fuzzy C-means clustering algorithm [18] to our similarity results. For evaluating the effectiveness of our proposed semantic similarity, we have performed extensive experiments. Our experiment results show that better clustering of semantically related pages with the lowest cluster duplication is achieved from the proposed similarity measure compared to other existing measures.

The rest of this paper is organized as follows. Section II reviews previous work on similarity measures and their use in creating page grouping through clustering. Section III introduces our proposed work with the new semantic similarity measure among the webpages. Section IV presents results from various experiments. Section V concludes the paper along with future work.

II. RELATED RESEARCH ACTIVITIES

Webpages often contain a number of distracting features and unnecessary objects such as advertisement, irrelevant video and/or audio, which may divert the user attention from the actual page content they are interested in. Indeed, an extension research works have been done to efficiently exclude the irrelevant entities and to successfully identify the keywords from the page. Generally, there are two types of keyword-extraction approaches [1]. One approach is domain-dependent based on supervised machine learning model, whereas other is domain-independent. Among all other keyword extraction approaches, TF_IDF (term frequency and inverse document frequency) weighting has been widely used [2]. Frank et al. [3] introduced an automatic keyword extraction algorithm (KEA) based on the TF_IDF, which was later enhanced by Kelleher et al. [4] through introducing Semantic Ratio (SR) feature. Typically, successful clustering of the web pages mainly depends on the similarity result of the extracted keywords of the pages. Prior research activities

focused on three different directions for discovering similarity among pages. They compared the text fragments as bags of words in vector space [5], using WorldNet [6], and also using Latent Semantic Analysis (*LSA*) [7]. However, Peng Qin [8] proposed a new method of page similarity based on WordNet hierarchy and the Directed Acyclic Graph (*DAG*). In addition to these measures, a number of similarity measures such as cosine similarity [9], Dice coefficient [10] and Jaccard coefficient [11] have been defined to compute alikeness. Later, Cilibrasi and Vitanyi [12] proposed a normalized google distance measure for computing the page similarity. However, some works also involved page-URL similarity for the alikeness computation [13]. For clustering similar web documents, there are two main approaches - hard clustering and soft clustering [14]. In K-means, presenting the hard Clustering, objects are partitioned into mutually exclusive clusters. However, in fuzzy C-means [18], presenting soft clustering, objects are partitioned into a number of clusters where each object is a member of every clusters with different degree of membership values. To assess the quality of page grouping, several cluster validation indices were introduced. Among them, the Davies Bouldin index [15], the Xie-Beni separation index [16], the Kwon index [17] are very well-known.

III. PROPOSED WORK

Our proposed work is divided into three major steps. At first, the webpages are preprocessed and then are represented by a vector of weighted keywords. Next, our proposed semantic similarity measure is used to determine the alikeness among the pages. Finally, a fuzzy clustering is applied to the similarity result for grouping the pages into a number of clusters.

A. Text Preprocessing

The target of the text processing is to represent each page by a set of weighted keywords. It is noted that a webpage can contain anything like texts, videos, audios and other multimedia objects. However, we consider only the text content of the page. The preprocessing consists of three sub-steps: (1) Stop-Word Removal, (2) Keyword Extraction, and (3) Assignment of Weights to Keywords.

(1) Stop-Word Removal

Stop words are basically high frequency words that are common in every text contents. These words include articles, prepositions, auxiliaries and modal verbs. Typically, they are not useful for representing the page individually. Besides, they carry a vast amount of unnecessary information which may mislead the mining operation. We filter out all stop words from the page content. Fig. 1 shows the stop word removal process.

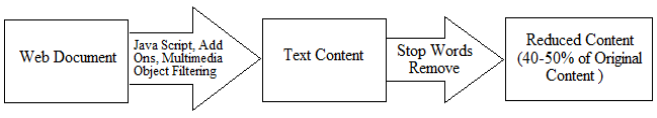

Figure 1. Unnecessary objects and stop word filtering.

(2) Keyword Extraction

Keywords are used to express the most important concepts of the page. True and meaningful representation of the page by a set of keywords relies entirely on the quality of keywords regarding their degrees of conceptual relatedness to the page. For effective keyword extraction, we again exclude some additional characters. Any character except English alphabet, numerical letter, double quotes, and single quotes is removed. From the remaining text, we extract only the noun and noun like words using the Parts Of Speech tagger (POS tagger) [19]. We consider only the noun and noun like words because in most cases they have the higher probability to be keywords. Besides, we reject all adjectives, verbs and other non-noun words.

(3) Assignment of Weights to Keywords

We assign weight to each keyword based on the following two factors:

(i) Frequency of the keyword in the page

(ii) Position of the keyword in the page

Generally, a frequent word (except stop words) is served as a keyword for the page. We use a lower bound for the frequency, where $1 > frequency\_threshold \leq F$, F is a positive numeral. In our work, any word with frequency greater than *frequency_threshold* is considered as a keyword and filter out all words with $frequency < frequency\_threshold$. Besides, the frequency, the position of a word in the page is also very vital. Here, the word position means the tag in which the word resides. Typically, the keywords present in the title tag have the highest relatedness to the pages. Therefore, these title words are the most important keywords to present the page and should have highest priority than any other keywords in the page. In addition, the keywords in the heading tags, representing the headline of a paragraph are also significant and should have higher priority. Furthermore, the keywords within the bold and italic tags are taking into account for higher priority assignment. Table I presents the tags and their respective weight using a scale of 1 to 10 employed in our approach for keyword weight assignment.

TABLE I. TAGS WITH THEIR POSITION WEIGHT

HTML Tags	Assigned Position Weight
<title>	10
<h1>	9
<h2>	8
<h3>	7
<h4>	6
<h5>	5
<h6>	4
<a>	3
<i>	2
<b>	2
<p>	1

A keyword can occur within multiple tags in a page, thereby having multiple position weights. For each keyword, we take a summation of all of its position weight while assigning weight. We retain only those keywords having *weight* > *weight_threshold* with $1 \geq weight\_threshold \leq W$, W is a positive numeral. Lastly, the keywords are rearranged according to their weight in a descending order so that the keyword most significantly representing the page takes place in the first position of the weighted keyword list. Our proposed approach for computing the keyword weight is presented in Algorithm I.

Algorithm I: Calculation of Keyword Weight W_i

Input: A set of words together with their respective frequency $\{(K_1, F_1), (K_2, F_2), \ldots, (K_i, F_i)\}$ extracted from the webpage P_t.

Output: A set of keywords with their respective weight $Keyword\_List_{p_t} = \{(K_1, W_1), (K_2, W_2), \ldots, (K_r, W_r)\}$.

```
for i = 1 to NumberOfWords do
    T_i ← FindTags(K_i, P_t)
    total_Weight ← 0
    for j = 1 to T_i.NumberOfTags do
        w_ij ← F_i × T_ij.position_Weight
        tota_Weight ← tota_Weight + w_ij
    end for
    W_i ← total_weight
    if W_i > weight_threshold then
        insert (K_r, W_r) into Keyword_List_pt.
    end if
end for
Rearrange Keyword_List_pt in a descending order
based on keyword weight.
return Keyword_List_pt.
```

B. *Proposed Semantic Similarity Measure*

Our proposed semantic similarity measure (*SSM*) takes into account both syntactic and semantic relatedness among the weighted keyword lists while computing similarity between two webpages. For similarity calculation between two keywords, we consider two situations: two keywords may be syntactically similar or both are different words but possess same meaning, i.e., semantically related words. It is noted that the syntactical similarity implicitly represents the semantic relationship between two keywords. Therefore, two pages have a similarity of one if they are represented by the same set of weighted keywords; otherwise they have varying similarity. For measuring the semantic relationship between two syntactically dissimilar keywords, we utilize the Synset of Wordnet [10]. In our proposed work, we verify the inclusion of one keyword to the Synset of other keyword. If they do, we consider them as a similar word; otherwise they are unrelated and distinct keywords. Equation (1) presents our proposed similarity measure between two webpages P_1 and P_2.

$$SSM(P_1, P_2) = \frac{\sum_{i=1}^{M \cap N} |P_1(K_i, W_i) + P_2(K_i, W_i)|}{\sum_{i=1}^{M} |P_1(K_i, W_i)| + \sum_{j=1}^{N} |P_2(K_j, W_j)|} \quad (1)$$

Here, $|M \cap N|$ denotes the size of semantic common keywords between P_1 and P_2. Our proposed similarity measure satisfies the following two properties:

- Identity: $SSM(P_1, P_2) = 1$
- Symmetry: $SSM(P_1, P_2) = SSM(P_2, P_1)$
- Uniqueness: $SSM(P_1, P_2) = 1$ means $P_1 = P_2$
- Positivity: $0 \leq SSM(P_2, P_1) \leq 1$

After similarity computation, we apply the fuzzy C-means clustering algorithm [18] for grouping the syntactically and/or semantically related pages.

IV. Experimental Results and Discussion

We performed a series of experiments to evaluate the effectiveness of proposed semantic similarity measure (*SSM*). To justify the performance of our *SSM*, we also carried out the same experiments with other similarity measures, namely, Cosine similarity (*CS*), Jaccard coefficient (*JC*), and the page-URL similarity (*US*) proposed in [13]. All experiments are accomplished on an Intel Core i3 Duo 3210 @3.20GHz desktop computer running on the Windows 7 operating system, with an installed RAM of 4.00 GB. For the experiments, we worked on 500 webpages collected from two popular English News-websites - *www.dailystar.net* and *www.bdnews24.com*. For evaluating the clustering of pages, we used three cluster validation indices – Davies-Bouldin index, Xie-Beni index, and Kwon index. All of these validation indices show better clustering of the webpages for the small index value.

A. *Performance Analysis with respect to Number of Clusters*

We verified three different index values by varying the number of clusters in the fuzzy C-means algorithm involving four different similarity computations [i.e., *SSM*, *CS*, *JC*, *US*] individually. Fig. 2, 3, and 4 present the values of Davies-Bouldin index (DBI), Xie-Beni index and Kwon index respectively for different number of clusters. We can observe that our *SSM* measure gives smaller index values for all three cases with varying cluster numbers. It is clear that after cluster number 10, the values of validation indices tend to grow up, representing the presence of large number of duplicate clusters. Table II and III present that the increasing number of clusters increases the percentage of cluster duplication. From the tables, we can clearly see that our SSN offers smaller percentage of cluster duplication in all cases. However, for the

cluster number 10, it shows the lowest duplication. As for both the cases (percentage of duplication and validation indices) SSM gives better result with the cluster number 10, we can conclude that for our experimental dataset the cluster number 10 is the optimum cluster number.

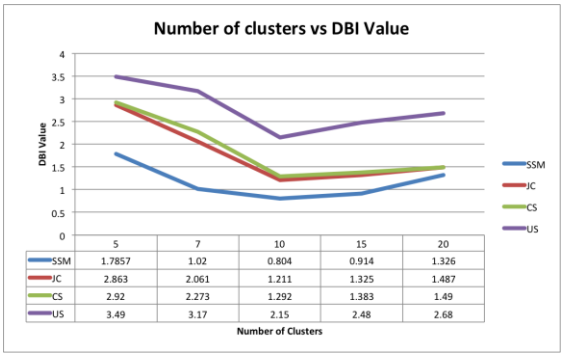

Figure 2. David- Bouldin index values for different number of cluster with different similarity approach.

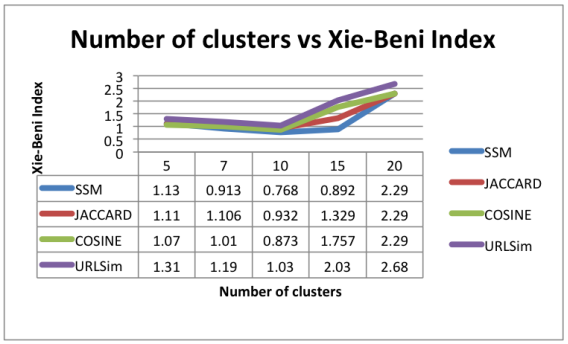

Figure 3. Xie-Beni index values for different number of cluster with different similarity approach.

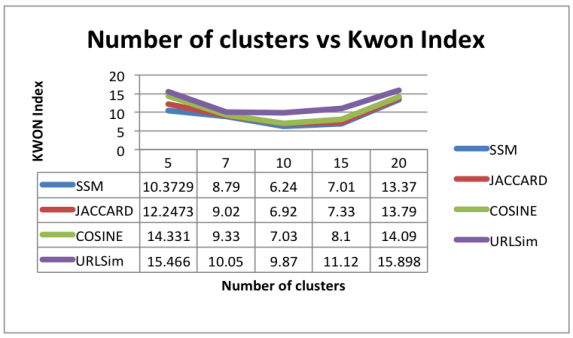

Figure 4. Kwon index values for different number of cluster with different similarity approach.

TABLE II. PERCENTAGE OF DUPLICATION IN THE FIRST 10 PAGES

Similarity Approaches	Number of Clusters			
	5	***10***	***15***	***20***
SSM	0%	0%	13%	15%
JC	0%	1%	19%	20%
CS	0%	1%	19%	20%
US	0%	2%	19%	21%

TABLE III. PERCENTAGEE OF DUPLICATION IN THE FIRST 20 PAGES

Similarity Approaches	Number of Clusters			
	5	***10***	***15***	***20***
SSM	0%	0%	14%	15%
JC	0%	1%	19%	20%
CS	0%	1%	19%	20%
US	0%	3%	22%	23%

B. Performance Analysis with respect to initial cluster membership selection

In fuzzy C-means clustering algorithm, the initial cluster-membership matrix can be generated in two ways. One way is to assign random membership value. However, other way is to use the membership value based on the similarity results between objects. Fig. 5 presents the value for David- Bouldin index with both these approaches for varying number of clusters. From this figure, we can observe that better cluster validation index can be achieved while using the similarity result obtained from our SSM for the membership matrix generation.

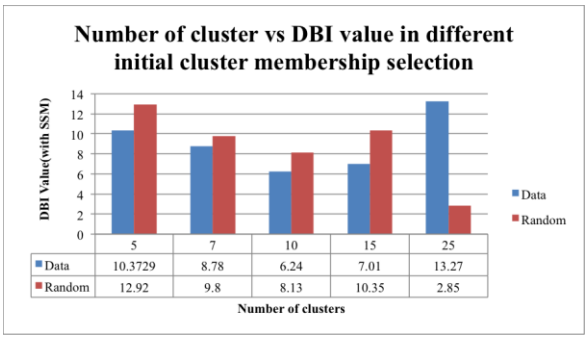

Figure 5. David- Bouldin index vs number of cluster with different initial cluster membership selection.

V. CONCLUSION

This paper presents a new semantic similarity measure for capturing the alikeness among the webpages. In our proposed approach, each page is represented by a vector of weighted keywords. Therefore, our similarity measure for two pages involves the similarity computation between the respective weighted keyword sets. Lastly, the pages are grouped by the fuzzy C-means clustering algorithm. Numerous experiments confirm that our similarity measure helps efficiently grouping the semantically related pages. Our experiments include performance comparison with other three popular similarity approaches. Our semantic similarity measure outperforms others by giving lower intra-cluster distance and higher inter-cluster distance. In near future, we aim to handle the text as well as other multimedia objects. We intend to utilize our offline semantically enriched page model using our proposed similarity to provide the online recommendations.

REFERENCES

[1] D. Chen, X. Li, J. Liu, and X. Chen, "Ranking-constrained keyword sequence extraction from web documents", Proceedings of the Twentieth Australasian Conference on Australasian Database, vol. 92, pp. 159-168, 2009.

[2] C. Salton and G.and Yang, "On the specification of term values in automatic indexing", Journal of Documentation, vol. 29(4), pp. 351 - 372, 1973.

[3] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning, "Domain-specific keyphrase extraction", Proceedings of the 16th International Joint Conference on Artificial Intelligence , pp. 668-673, 1999.

[4] D. Kelleher and S. Luz, "Automatic hypertext keyphrase detection", Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 1608-1609, 2005.

[5] R. A. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1999.

[6] C. Fellbaum, "WordNet: An Electronic Lexical Database", The MIT Press, May 1998.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis", Journal of the American Society for Information Science, vol.41(6), pp.391-407, 1990.

[8] P. Qin, Z. Lu, Y. Yan, and F. Wu, "A New Measure of Word Semantic Similarity based on WordNet Hierarchy and DAG Theory", International Conference on Web Information Systems and Mining, pp. 181-185, China, 2009.

[9] B. Hajian and T. White, "Measuring Semantic Similarity using a Multi-Tree Model", Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2011), pp. 7-14, Barcelona, Spain, July 2011.

[10] C. D. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval", Cambridge University Press, Cambridge, UK, 2008.

[11] L. Hamers, Y Hemycker, G. Herweyears, M. Janssen, R. Rousseau, and A. Vanhoutte, "Similarity measure in scientometric research: The Jaccard Index versus Salton's Cosine formula", Information Processing & Management, vol. 25(3), pp. 315–318, 1989.

[12] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google Similarity Distance", IEEE Transactions on Knowledge And Data Engineering, vol. 19(3), pp. 370-383, March 2007.

[13] S. Kabir, S. P. Mudur, and N. Shiri, "New similarity measures for capturing browsing interests of users into web usage profiles", Proceedings of AAAI Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2012), pp. 18-25, Toronto, Canada, July 2012.

[14] S. Ghosh and S. K. Dubey, "Comparative analysis of K-means and Fuzzy C-means algorithms", International Journal of Advanced Computer Science & Applications, vol. 4(4), pp. 35-38, 2013.

[15] D. L. Davies and D. W. Bouldin, "Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1(2), pp. 224-227, 1979.

[16] X. L. Xie and G. Beni, "A Validity measure for Fuzzy Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13(8), pp. 841-847 , 1991.

[17] S. H. Kwon, "Cluster validity index for fuzzy clustering", Eletronics Letters, vol. 34(22), pp. 2176-2177, 1998.

[18] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact, well separated clusters", Journal of Cybernetics, vol. 3(3), pp. 95-104 , 1974.

[19] S. Goldwater and T. L. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging", Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 744-751, Prague, Czech Republic: Association for Comp. Linguistics, 2007