IJCSE International Journal of Computer Sciences and Engineering Open Access

Short Communication Volume-2, Issue-8 E-ISSN: 2347-2693

Weighted Word Affinity Graph for Betterment of Spatial Information Descriptors

Dr.Poonam Yadav[1]

D.A.V College of Engineering. & Technology, India

www.ijcseonline.org

Received: 16/07/ 2014 Revised: 28/07/ 2014 Accepted: 24 /08/ 2014 Published: 31 /08/ 2014

Abstract**—** Document analysis/ retrieval system plays crucial role to strengthen any information retrieval system. There are various processing stages associated with a document analysis system, such as feature extraction stage, semantic representation stage, dimensionality reduction stage and similarity measure stage. Researchers are contributing well in every stage to improve the performance of the document analysis system. This short paper considers word affinity graph/ matrix for further improvement so that semantic representation can be given more precisely. This is accomplished by incorporating weight component in the word affinity matrix to provide significance for degree of distribution. Theoretical study on both word affinity matrix and weighted word affinity matrix shows the significance offering by them on widely distributed document terms.

Keywords**—**Information retrieval, Document analysis, affinity graph, weighted word affinity graph, semantic

I. Introduction

Nowadays, usage of electronic documents are highly appreciated because of their compactness, easy to explore and retrieve, fast documents transfer, etc. Hence, all the paper documents are being converted to electronic documents [1]. On the other hand, data analysts and investigators attempt to explore huge databases to solve problems and making decisions for betterment of their respective field [2]. However, statistical and computational analyses require a huge collection of text documents with numerical or structured values. This leads the information retrieval system to process slowly and sequentially as well [2]. This system extracts the documents from huge databases based on the requirements of the users [7] [8].

In – depth document analysis play crucial role for such information retrieval system. However, the documents have different words to describe a concept and alternative words with similar meaning. This poses a great challenge for conventional information retrieval system. The challenge gets increased further, when the query becomes too long [3]. Hence, semantic concept has been introduced and now it has found great attention towards research [9].

II. Document analysis/ retrieval system

A general layout of a document analysis and retrieval system is given in Fig. 1. The training phase constitutes a process of constructing a feature library in which the training documents are well – treated and features are stored in structured manner. The first process extracts features, may be local features or global features or both, from the subjected documents. The extracted features are subjected to understand the semantic concepts behind them. The semantic representation of the extracted features may be high dimensional, and sometimes multi – dimensional. Hence, dimensionality reduction methods cross the path and play its role. Once the semantic representation of extracted features is obtained in low dimension, they are structured and stored in feature library. Given a test document, the similar documents can be retrieved from the database by performing all the aforesaid processes of the test document followed by measuring the similarity of it with the feature library contents.

III. Motivation

Numerous research works have been carried out in the literature to develop such a retrieval system by contributing towards betterment of each processing stages. For instance, vector space model (VSM) [11] has been introduced to use tf – idf weighting scheme and to generate a feature vector at constant length, instead of varying length [4]. In another case, Latent Semantic Indexing (LSI) [11] [12] has been used for dimensionality reduction. However, Principle Component Analysis (PCA) [14] has replaced LSI over the period by solving the dimensionality reduction problem as an eigenvalue problem [5]. All the components have outperformed well as per the literary report [10]. However, by improving each component, the overall retrieval accuracy and precision can be significantly improved further.

This short paper presents a theoretical outline on improved semantic representation for the betterment of document analysis system. We have presented our thought process of using weights with word affinity graph based on the frequency of occurrence and co-occurrence measures.

The rest of this short paper is organized as follows. Section 4 briefs the word affinity graph in a document analysis system

Corresponding Author: *Dr.Poonam Yadav,*
poonam.y2002@gmail.com

and Section 5 details the weighted word affinity graph. Section 6 conducts a theoretical investigation to understand the performance variation and Section 7 concludes the paper.

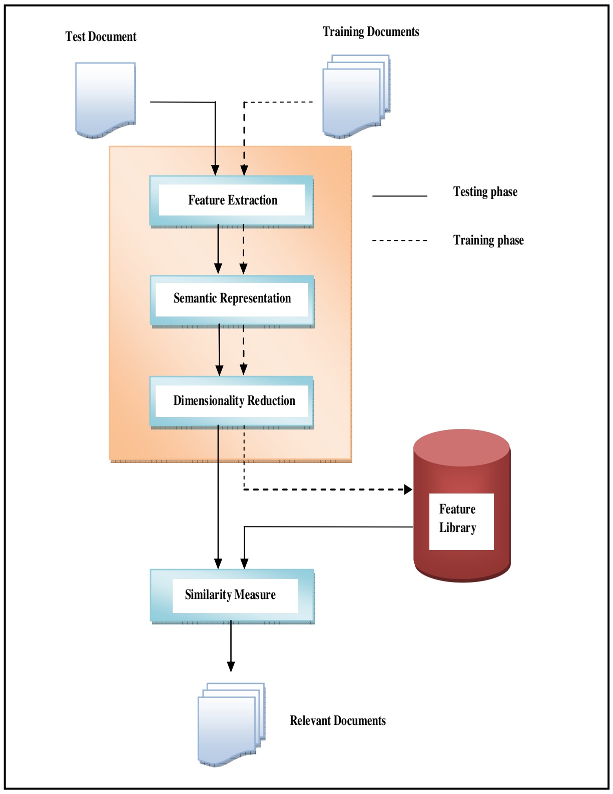

Fig. 1. General architecture of a document analysis/ retrieval system

IV. Word Affinity Graph

A word affinity graph is a document representation given by frequency of term co-occurrence in a document section [6]. Consider two words in a paragraph, say a and b. The characteristic of individual and joint occurrence of these terms constitute word affinity graph. In contrast to the name, the word affinity graph is often represented in matrix format G, where the element of the matrix can be represented as

$$g_{m,n} = \begin{cases} \sum_{i=1}^{N} f_i(m); m = n \\ |D_1|; otherwise \end{cases} \qquad (1)$$

where, N represents number of documents, m and n represent a and b, respectively and $|D_1|$ represents cardinality of the set D_1, subjected to the constraint, $D_1 \subseteq D : d_i \in D_1; when\ a,b \in d_i$. Here, D is a set of documents represented as $\{d_1, d_2, \ldots, d_N\}$; d_i refers to i^{th} document. The generated matrix can also be called as word affinity matrix.

V. Weighted Word Affinity Graph

This work recommends considering joint probability distribution function to construct the weighted word affinity matrix G'. Hence, the modified formula can be given as

$$g'_{m,n} = \begin{cases} \sum_{i=1}^{N} f_i(m); m = n \\ \dfrac{|D_1|}{idf(m) \cdot idf(n) + \alpha}; otherwise \end{cases} \qquad (2)$$

where, $idf(\bullet)$ represents inverse document frequency of the respective word represents a small scaling factor (usually set as 0.1) to avoid infinite value. This acts as the weight to impose the significance of the co-occurrence factor given in the numerator. In other words, a matrix element becomes high, when the co-occurrence and individual document frequency are almost similar. Hence, the weighted word affinity matrix describes a document based on the complete spatial distribution of the feature, whereas the word affinity matrix may consider the term distribution partially. It has also to be noted that *tf* (term frequency) and *idf* are the most prevalently used term weighting measures as it quenches the needs to accomplish information retrieval accuracy [13].

VI. Peformance Investigation

Let us consider two words 'hi' and 'how' and there are 10 documents in the database. The term frequencies of these two terms are given in Table I. Based on Table I, the constituents for word affinity matrix and weighted word affinity matrix can be tabulated as given in Table II and V, respectively. Table III gives both the resultant word affinity matrix and weighted word affinity matrix, where diagonal elements are same and non – diagonal elements are different. The non – diagonal elements of weighted word affinity matrix are lesser than those in the word affinity matrix. This interprets that these two words are not well distributed throughout all the documents.

Consider the sample example but with different frequency distribution as given in Table IV and so the constituents table can be constructed as Table V. The frequency distribution is different, but for convenient comparison, the term frequency is maintained same.

Table VI shows the constructed word affinity matrix and weighted word affinity matrix, where the non – diagonal elements of the weighted word affinity matrix are ten times greater than that of the word affinity matrix. This is achieved because of the sparse distribution of the terms throughout all the documents and so higher importance is given here compared to the previous example. Hence by exploiting weighted word affinity matrix, the feature descriptors can be more informative by providing significance to the distribution ratio.

TABLE I: TERM FREQUENCIES OF THE WORDS '*HI*' AND '*HELLO*' WEAK TERM SPATIAL DISTRIBUTION

Document	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
"*Hi*"	3	4	0	2	0	0	1	0	4	0
"*How*"	2	0	0	1	0	3	0	0	0	5

TABLE II: CONSTITUENTS TO BUILD WORD AFFINITY GRAPH AND WEIGHTED WORD AFFINITY GRAPH WEAK TERM SPATIAL DISTRIBUTION

Document	**Word 1 (*Hi*)**		**Word 2 (*How*)**		**Co – occurrence count**
	Frequency	***Availability Index***	***Frequency***	***Availability Index***	
1	3	1	2	1	1
2	4	1	0	0	0
3	0	0	0	0	0
4	2	1	1	1	1
5	0	0	0	0	0
6	0	0	3	1	0
7	1	1	0	0	0
8	0	0	0	0	0
9	4	1	0	0	0
10	0	0	5	1	0
	14 (Term frequency)	5 (total availability index)	11 (Term frequency)	4 (total availability index)	2 (joint co-occurence)
		-1 (idf)		-1.32 (idf)	

TABLE III: COMPARISON BETWEEN WORD AFFINITY MATRIX AND WEIGHTED WORD AFFINITY MATRIX ON DOCUMENTS HAVING WEAK SPATIAL DISTRIBUTION

Word Affinity Matrix	**Weighted Word Affinity Matrix**
$\begin{bmatrix} 14 & 2 \\ 2 & 11 \end{bmatrix}$	$\begin{bmatrix} 14 & 1.41 \\ 1.41 & 11 \end{bmatrix}$

TABLE IV: TERM FREQUENCIES OF THE WORDS '*HI*' AND '*HELLO*' WITH STRONG TERM SPATIAL DISTRIBUTION

Document	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
"*Hi*"	2	2	2	2	1	1	1	1	1	1
"*How*"	1	2	1	1	1	1	1	1	1	1

TABLE V: CONSTITUENTS TO BUILD WORD AFFINITY GRAPH AND WEIGHTED WORD AFFINITY GRAPH WEAK TERM SPATIAL DISTRIBUTION

Document	**Word 1 (*Hi*)**		**Word 2 (*How*)**		**Co – occurrence count**
	Frequency	***Availability Index***	***Frequency***	***Availability Index***	
1	2	1	1	1	1
2	2	1	2	1	1
3	2	1	1	1	1
4	2	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1
10	1	1	1	1	1
	14 (Term frequency)	10 (total availability index)	11 (Term frequency)	10 (total availability index)	10 (joint co-occurence)
		0 (idf)		0 (idf)	

TABLE III: COMPARISON BETWEEN WORD AFFINITY MATRIX AND WEIGHTED WORD AFFINITY MATRIX ON DOCUMENTS HAVING STRONG SPATIAL DISTRIBUTION

Word Affinity Matrix	**Weighted Word Affinity Matrix**
$\begin{bmatrix} 14 & 10 \\ 10 & 11 \end{bmatrix}$	$\begin{bmatrix} 14 & 100 \\ 100 & 11 \end{bmatrix}$

VII. CONCLUSION

Here, we discussed the necessity of document analysis system and appropriate selection of processing stages and their techniques. Further, we introduced a weighted word affinity matrix for providing better feature descriptors for the further stages of the document analysis system. This weighted word affinity matrix claimed that it provide higher significance to well distributed terms, when compared with conventional word affinity matrix. An exemplary description was presented to support the claim through which ten times higher significance was observed for fully distributed terms in weighted word affinity matrix over the conventional word affinity matrix.

REFERENCES

[1] Song Mao, Azriel Rosenfeld, Tapas Kanungo, "Document structure analysis algorithms: a literature survey", DRR 2003, **2003**, p.p. **197-207**

[2] Carsten Gorg, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, Haesun Park, Member, and John Stasko, "Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw", IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 10, **2013**, p.p. **1646 – 1663**

[3] Jinxi Xu Amherst, W. Bruce Croft, "Query expansion using local and global document analysis", Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, **1996**, p.p. **4-11**

[4] G. Salton, M. McGill, Eds. "Introduction to Modern Information Retrieval", New York: McGraw-Hill, **1983**.

[5] S. Deerwester and S. Dumais, "Indexing by latent semantic analysis," J. Amer. Soc. Inf. Sci., vol. 41, no. 6, **1990**, pp. **391–407**.

[6] Haijun Zhang, John K. L. Ho, Q. M. Jonathan Wu, Senior Member, IEEE, and Yunming Ye, "Multidimensional Latent Semantic Analysis Using Term Spatial Information", IEEE Transactions on Cybernetics, Vol. 43, No. 6, **2013**, p.p. **1625-1640**

[7] W. B. Frakes and R. Baeza-Yates, "Information Retrieval: Data Structures and Algorithms", Prentice-Hall, Englewood Cliffs, NJ, **1992.**

[8] Antoniol, G. ; Canfora, G. ; Casazza, G. ; De Lucia, A; "Information retrieval models for recovering traceability links between code and documentation", Proceedings of International Conference on Software Maintenance, **2000**, p.p. **40-49**

[9] Yu-Gang Jiang ; Yang, J. ; Chong-Wah Ngo ; Hauptmann, A.G.; "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study", IEEE Transactions on Multimedia, Vol. 12, No. 1, **Jan. 2010**, p.p. **42 – 53.**

[10] Eaddy, M. ; Antoniol, G. ; Gueheneuc, Y.-G., "CERBERUS: Tracing Requirements to Source Code Using Information

Retrieval, Dynamic Analysis, and Program Analysis", 16th IEEE International Conference on Program Comprehension (ICPC 2008), 10-13 June **2008**, p.p. **53 - 62**

[11] G. Antoniol, G. Canfora, G. Casazza, A. De Lucia, E. Merlo, "Recovering Traceability Links between Code and Documentation," IEEE Transactions on Software Engineering, Vol .28, No. 10, **2002**, p.p.**970–983**

[12] D. Poshyvanyk, Y.-G. Guéhéneuc, A. Marcus, G. Antoniol, V. Rajlich, "Feature Location Using Probabilistic Ranking of Methods Based on Execution Scenarios and Information Retrieval," IEEE Transactions on Software Engineering, Vol. 33, No. 6, **2007**, p.p.**420–432.**

[13] Akiko Aizawa, "An information-theoretic perspective of tf–idf measures", Information Processing and Management, Vol. 39, **2003**, p.p. **45–65**

[14] Wray Buntine and Aleks Jakulin, "Applying discrete PCA in data analysis", Proceedings of the 20th conference on Uncertainty in artificial intelligence, **2004**, p.p. **59-66**

Dr. PoonamYadav obtained B.Tech in Computer Science &Engg. fromKurukshetra University Kurukshetra and M.Tech in Information Technology from Guru Govind Singh Indraprastha University in 2002 and 2007 respectively. She had Awarded Ph.D inComputer Science& Engg. fromNIMS University, Jaipur. She is currently working as Principal in D.A.V College of Engg. & Technology, Kanina (Mohindergarh). Her research interests include Information Retrieval, Web based retrieval and Semantic Web etc. Dr. PoonamYadav is a life time member of Indian Society for Technical Education and her email id is poonam.ir@gmail.com.