Enliven Archive
Intensifying Thoughts

Review Article www.enlivenarchive.org Enliven: Journal of Genetic, Molecular and Cellular Biology

ISSN: 2379-5700

Analysis of Improved Semantic Approaches Based on Topic Maps

Vincent Marcov

Department of Pediatrics and Genetics, Ain Shams University, Nour Mosque, 38 Abbasia Next, El-Mohamady, Al Waili, Egypt

*Corresponding author: Vincent Marcov, Department of Pediatrics and Genetics, Ain Shams University, Nour Mosque، 38 Abbasia Next، El-Mohamady, Al Waili, Egypt, E-mail: vincentmarcov@protonmail.com

Received Date: 05th March 2018
Accepted Date: 21st April 2018
Published Date: 01st May 2018

Citation: Marcov V (2018) Analysis of Improved Semantic Approaches Based on Topic Maps. Enliven: J Genet Mol Cell Biol 5(2): 003.

The researchers argue that while performing document clustering, calculating the similarity measure between a pair of documents is a significant computational burden. Similarity measure involves determining the extent of similarity between a couple of reports then they are assigned a real number (0-1) [1]. A value of one shows that the documents are essentially the same while a value of zero signifies that the documents are entirely disparate. Conventionally, the vector-based models are usually relied to compute similarity between a pair of documents [2]. The models represent diverse characteristics of the documents; nonetheless, they cannot account for the document's semantics [3]. Published documents usually have contexts in addition semantically related words are mostly used explain such contexts [4]. Fascinated by such a fact, different scientists use text annotation via external thesauruses such as lexical database (Word Net) to investigate semantic-based similarity measures [5].

In this study, the researchers used documents which were represented in Topic Maps to describe a semantic similarity. A Topic Map is a software popularly used to encode knowledge and connect the relevant knowledge [6]. The topic maps using the similarity measure are meaningfully superior as compared to current techniques used to calculate document similarity. The maps increasingly turning to be a practice standard for knowledge representation [7]. The coded knowledge is relied to convert the document into a map; besides, a relationship between the common patterns is used to represent the similarity between a pair of documents [8]. The text mining datasets are subjected to experimental studies which ultimately show that the new similarity measure is very successful in performing text clustering in comparison to the mostly utilized similarity measures [9]. Strehl et al. were among the first to investigate the effect of similarity measure with regards to performing clustering [10]. The researchers utilized the YAHOO datasets to make comparison of diverse similarity measures such as Extended Jaccard, Pearson correlation, Cosine, and Euclidean. Furthermore, they used diverse clustering algorithms such as weighted-graph partitioning, hyper-graph partitioning, self-organizing features map, as well as Generalized K-mean [11]. The experiment revealed that Cosine and Extended Jaccard were closer to outcomes related to categorizations done by human.

The precision level of clustering is mainly determined by the efficiency of the defined similarity measure [12]. It becomes relatively easy to compute the function of similarity measure; besides, they must be easily explained and be able to subtly capture the correlation between the documents [13]. With regards to document clustering, researchers have proposed numerous similarity measures for instance, the most popular one is the cosine similarity measure and the Euclidean measure that relies on the distance metric to calculate documents similarity [14]. The two measures usually disregard the words or terms' semantics as contained in the document. In this case, the researchers adopted the Word Net-based semantic similarity to compute the semantic similarity [15]. They computed similarity which is based on the common concepts hierarchy by initially extracting semantic words from the documents and then developed a semantic class hierarchy of the terms.

The researchers performed clustering using a hierarchical agglomerative clustering technique to assess the efficiency of the recommended similarity measure [16]. Ultimately, the researchers concluded that clustering of documents collection could effectively be done using the topic based similarity measure since it generated more rational clustering in comparison to the human categorized structures [17]. Inspection of the outcomes on purity proves that with the exception of NEWS20, the KLD similarity emerged the second best in all sets of data [18]. The reason for performing poorly in the NEWS20 dataset is the fact that the latter is characterized by categories which are closely related as well as the collection contains documents with shorter lengths [19]. Alternatively, it is evident that in all datasets the cosine similarity performed relatively well; conventionally, it was a selection to perform clustering the collection of documents.

The outcome on entropy additionally validates the researcher's findings that the topic map is an exceptional technique to perform representation of documents for clustering [20]. Choosing the general distribution of the collection as well as making a decision about the precise cluster of the document can be enhanced using the representation as evidently shown in the entropy results on the NEWS20 dataset [21]. Furthermore, the researchers argue that numerous approaches can be obtained for future research or studies for instance, there

might be need to compare the performance of clustering as well as the documents representation [22]. Additionally, there can be need to examine the impact of such measures by suing diverse type of clustering algorithms.

Table 3: Entropy with different similarity measures

Dataset	Euclidean	Cosine	Jaccard	KLD	TM-sim
NEWS20	0.95	0.51	0.52	0.54	**0.35**
Reuters	0.6	0.27	0.33	0.26	**0.21**
Webkb	0.9	0.66	0.68	0.44	**0.4**
Classic	0.78	0.29	0.2	0.3	**0.2**
OSHUMED	0.43	0.32	0.26	0.21	**0.2**

The same result in graphical form.

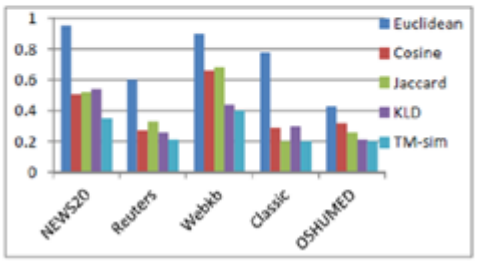

Figure 1: Graphical representation of entropy [20]

References

1. Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5: 001.
2. Rafi Muhammad, Mohammad Shahid Shaikh (2013) An Improved Semantic Similarity Measure aor Document Clustering Based on Topic Maps. Computer Science Department NU-FAST 4087: 1-5.
3. Lin Jianhua (1991) Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory 37: 145-151.
4. Huang Anna (2008) Similarity Measures for Text Document Clustering. In Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008) 37: 49-56.
5. Larsen, Bjornar, Chinatsu Aone (1999) Fast and Effective Text Mining Using Linear-Time Document Clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 6: 16-22.
6. Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4: 001.
7. Pedersen Ted, Siddharth Patwardhan, Jason Michelizzi (2004) Word Net: Similarity: Measuring the Relatedness of Concepts. In Demonstration papers at HLT-NAACL 38-41.
8. Peter Willett (1988) Recent Trends in Hierarchic Document Clustering: A Critical Review. Information Processing &Management 24: 577-597.
9. Diana McCarthy (2006) Relating Word Net Senses for Word Sense Disambiguation. In Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together.
10. Shehata Shady (2009) A Word Net-Based Semantic Model for Enhancing Text Clustering. In Data Mining Workshops, ICDMW'09 IEEE 477-482.
11. Shehata Shady, Fakhri Karray, Mohamed S Kamel (2010) An Efficient Model for Enhancing Text Categorization Using Sentence Semantics. Computational Intelligence 26: 215-231.
12. Alexander Strehl, Joydeep Ghosh, Raymond Mooney (2000) Impact of Similarity Measures on Web-Page Clustering. In Workshop on Artificial Intelligence for Web Search (AAAI 2000) 5: 64.
13. Shehata Shady, Fakhri Karray, Mohamed Kamel (2008) Enhancing Text Categorization Using Sentence Semantics. In International Conference on Advanced Data Mining and Applications, Springer 87-98.
14. Chun-Ling Chen, Frank SC Tseng, Tyne Liang (2010) Mining Fuzzy Frequent Itemsets for Hierarchical Document Clustering. Information processing & management 46: 193-211.
15. Pepper Steve, and G. Moore (2010) Topic Maps. Encyclopedia of Library and Information sciences 1: 5247-5259.
16. Malik Hassan H, John R Kender (2006) High Quality, Efficient Hierarchical Document Clustering Using Closed Interesting Itemsets. In Data Mining ICDM'06. Sixth International Conference 7: 991-996.
17. Kawasaki Saori, Ngoc Binh, Tu Bao (2000) Hierarchical Document Clustering Based On Tolerance Rough Set Model. In European Conference on Principles of Data Mining and Knowledge Discovery, Springer 458-463.
18. Oren Zamir, Oren Etzioni, Omid Madani, Richard M Karp (1997) Fast and Intuitive Clustering of Web Documents. In KDD 2: 287-290.
19. Lutz Maicher, Lars Marius Garshol (2008) Scaling Topic Maps: Third International Conference on Topic Map Research and Applications. TMRA, Springer.
20. Wei Xu, Xin Liu, Yihong Gong (2003) Document Clustering Based on Non-Negative Matrix Factorization. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2: 267-273.
21. Wen-tau Yih, Christopher Meek (2007) Improving Similarity Measures for Short Segments of Text. In AAAI 7: 1489-1494.
22. Benjamin CM Fung, Ke Wang, Martin Ester (2003) Hierarchical Document Clustering Using Frequent Item Sets. In Proceedings of the 2003 SIAM international conference on data mining 59-70. Society for Industrial and Applied Mathematics.
23. Yong Wang, Julia Hodges (2006) Document Clustering with Semantic Analysis. Proceedings of the 39th Annual Hawaii International Conference, IEEE 3:54c-54c.
24. Yong Wang, Julia E Hodges (2005) Document Clustering Using Compound Words. In IC-AI 2: 307-313.
25. Swathi M (2017) Drug Prediction of Cancer Genes Using SVM. Enliven: Pharmacovigilance and Drug Safety 4: 001