Webpage Recommendation System Based on the Social Media Semantic Details of the Website

Dr. R.Rooba [1], Mr. S.Muruganantham [2]

[1,2] Assistant Professor ,Kongu Arts and Science College (Autonomous), Erode, Tamil Nadu, India,
[1]rrooba@gmail.com
[2]muruganandham.s@gmail.com

Article History:Received:11 november 2020; Accepted: 27 December 2020; Published online: 5 April 2021

Abstract: The web page recommendation is generated by using the navigational history from web server log files. Semantic Variable Length Markov Chain Model (SVLMC) is a web page recommendation system used to generate recommendation by combining a higher order Markov model with rich semantic data. The problem of state space complexity and time complexity in SVLMC was resolved by Semantic Variable Length confidence pruned Markov Chain Model (SVLCPMC) and Support vector machine based SVLCPMC (SSVLCPMC) methods respectively. The recommendation accuracy was further improved by quickest change detection using Kullback-Leibler Divergence method. In this paper, socio semantic information is included with the similarity score which improves the recommendation accuracy. The social information from the social websites such as twitter is considered for web page recommendation. Initially number of web pages is collected and the similarity between web pages is computed by comparing their semantic information. The term frequency and inverse document frequency (tf-idf) is used to produce a composite weight, the most important terms in the web pages are extracted. Then the Pointwise Mutual Information (PMI) between the most important terms and the terms in the twitter dataset are calculated. The PMI metric measures the closeness between the twitter terms and the most important terms in the web pages. Then this measure is added with the similarity score matrix to provide the socio semantic search information for recommendation generation. The experimental results show that the proposed method has better performance in terms of prediction accuracy, precision, F1 measure, R measure and coverage.

Keywords: Web page recommendation, socio semantic information, point wise mutual information, recommendation generation.

1 Introduction

World Wide Web (WWW) has become the most popular way of communicating, retrieving and disseminating information. The number of web pages keeps growing very rapidly. Web Page recommendation (Bhavsar, M., & Chavan, M. P. 2014) is developing popular websites and it links to related or similar stories, books or most visited pages at websites.

Web page recommendation system (Waykule, V., & Gupta, S. S. 2014) can be utilized to find out the personalized web service by suggesting the pages that are likely to be accessed in future. Web page recommendation system understands the user navigation pattern by exploiting the web usage mining provides personalization based on the results of mining. For the prediction of user's next link of choice and for pre-fetching links, Markov models were more popularly used (Shirgave, S. et al. 2014). But it has issues like high state space complexity, low coverage and low prediction accuracy. These issues are overcome by SVLMC model. But it doesn't consider the out link of the state that also influences the accuracy of next link prediction. This was overcome by using Confidence-Pruned Markov Model (CPMM) in SVLMC that considers both out-links and in-links of the state during pruning process estimation to rank the web pages.

In this paper, recommendation accuracy is further improved by including socio semantic information with the similarity score. A number of web pages are collected and then based on semantic information the similarity between two web pages are calculated. A composite weight for each semantic metadata in the web page is generated by using term frequency and inverse document frequency (tf-idf) and it returns the most important terms in the web pages. Then the closeness between the most important terms and the twitter terms are calculated by using Point Wise Mutual Information (PMI) and it is added with the similarity score matrix which provides socio semantic information for recommendation generation. It improves the recommendation accuracy.

*Corresponding author: Dr. R.Rooba
Assistant Professor ,Kongu Arts and Science College (Autonomous), Erode, Tamil Nadu, India,
rrooba@gmail.com

2 Literature Survey

A novel web page recommendation system method (Nguyen, T. T. S. et al. 2014) was proposed for efficient web page recommendation. It was achieved through semantic enhancement where the domain knowledge and the web usage of a website were combined. In order to represent the domain knowledge of a website, this model utilized an ontology which represented the domain knowledge and another model represented the web pages, domain terms and the relations between them by using automatically generated semantic network.

A hybrid approach (Wen, H. et al. 2012) was proposed for personalized recommendation of news on the web. It provided web users with an autonomous tool which was able to reduce tedious and repetitive web surfing. Initially in the proposed hybrid approach the weights of terms in web pages were calculated and based on that web pages were classified. Through analyzing the user's navigational history a user's interest and preference models were generated. Based on the user's model, content of web pages and preference models, the recommender system suggested news web pages to the user who was likely interested in the related topics. But still this approach has scalability issues.

WebPUM is a web page recommendation system was proposed (Jalali, M. et al. 2010) for prediction of user future movements. It was based on graph partitioning to model user navigation patterns during the mining phase. But this system has poor quality of recommendations.

A hybrid web page recommendation system (Abrishami, S. et al. 2012) was designed for web page recommendation. It was based on combining semantic information with web usage mining and page clustering based on semantic similarity. Because web pages were seen as ontology individuals, frequent navigational patterns were in the form of ontology instances instead of page clustering and page addresses was done using semantic similarity. The result was utilized to produce web page recommendation to users. The recommender system utilized in the hybrid web page recommendation system based on page clustering and semantic patterns which created a list of appropriate recommendations. Still the hybrid web page recommendation system has poor precision value.

A web page recommendation system (Rizvi, N. T. S. H., & Keole, R. R. 2015) was proposed in information retrieval using domain knowledge and web usage mining. It was a desktop search utility which found term patterns in web query data by using web usage mining process. This was utilized for the prediction of possible next pages in the browsing sessions. The proposed web page recommendation system consists of four phases are extracting pages from Google, creating term patterns, probability calculation of term patterns in name of web pages and recommending new list of web pages based on term frequencies. The sequential web term patterns were extracted by using Sequential Pattern Mining (SPM). The mined sequential patterns were stored in the software folder and then these used for matching and generating web links for online recommendations. But this system failed to concentrate on web usage knowledge base update, multi-site, domain knowledge discovery, web usage data and representation for web page recommendation.

A semantically enriched web usage based recommendation model (Ramesh, C. et al. 2011) combined semantic information with web usage mining process. The sequential pattern mining technique was applied in the semantic space to find out the frequent sequential patterns. In the form of ontology instances the frequent navigational patterns were extracted. The resultant semantic patterns were utilized to generate web page recommendations to the user. The extracted frequent patterns reflect the semantic relatedness between the visited web pages. The discovered semantic rich sequential association rules from the core knowledge of the recommendation of the proposed semantically enriched web usage based recommendation model.

A framework by (Rizvi, N. S., & Keole, R. R. 2015) were proposed to signify the domain knowledge of a web page. One of the models is an ontology based model which can be built semi-automatically called as DomainOntoWP and another model is a semantic network of web pages which can also be built automatically called as TermNetWP. In this framework, a novel method was used which provided a web page recommendation through semantic enhancement by combining the domain and web usage knowledge of a web page.

The model presented by (Nguyen, T. T. S. et al 2010) represents the non-taxonomic visiting relationship present among the visited pages. The result of this model was an ontology style document that provided the web usage knowledge to be machine understandable and sharable in recommender systems of semantic web applications. The presented model signified the weighted graph created from the raw web logs through the ontology language OWL. It captures the non taxonomic relations between the visited pages generated from web usage mining that supports enrichment of domain ontology and semantic Web recommendation. However, the presented model might be adapted to develop some Web 3.0 applications to handle user's decision making of web browsing process.

3 Proposed Methodology

In this section, the proposed method for socio semantic information based recommendation generation is described in detail. The Support vector machine based Semantic Variable Length Confidence Pruned Markov Chain (SSVLCPMC) model is determined the transition probability between web pages and the Kullback-Leibler Divergence method is used for quickest change detection. Initially the twitter data are collected and the most important terms in the web pages are collected based on the composite weight of each semantic metadata. The PMI is calculated between the most important terms and the terms in the tweets are calculated and it is called as twitter score. Finally it is added with similarity score which improves the recommendation accuracy.

3.1 Similarity Score Calculation

Initially, a semantically enriched transition probability matrix is obtained by using SSVLCPMC model. Then the similarity between any two web pages is determined by comparing semantic information between those two web pages. A similarity score between two web pages is incremented by one for every two semantic items that are identical. The similarity score is calculated based on Term Frequency-Inverse Document Frequency (TF-IDF). Here, a web page is represented by an n-dimensional vector h, where x is the total number of semantic metadata items in the web site. The vector for each web page is represented as $h = \{y(f_1, n), y(f_2, n), \dots y(f_p, n)\}$, where $y(f_j, n)$, for $1 \leq j \leq p$, is the weight of the j-th semantic metadata item f_j in the page n. The TF-IDF is utilized to produce a composite weight for each semantic metadata in a web page. The Term Frequency (TF) of a semantic metadata item f_j is given as follows:

$$TF_{f_j,n} = \frac{frequency\ of\ semnatic\ metadata\ item\ f_j\ in\ web\ page\ n}{N_n}$$

where, N_n is the total number of semantic meta data items in the web page n. The inverse document frequency (IDF) of semantic metadata item f_j is given as follows:

$$IDF_{f_j} = \log \frac{N}{df_j}$$

where N is the total number of web pages in web site and df_j is the document frequency of the semantic metadata item f_j i.e., the number of web pages in the web site containing the semantic metadata item f_j. The weight of the semantic metadata item in page is given by the TF-IDF weighting schema is given as follows:

$$y(f_1, n) = TF_{f_j,n} \times IDF_{f_j}$$

The similarity between two web pages is calculated by making use of Tanimoto coefficient and it is computed by using following equation:

$$Similarity\ Score(N_1, N_2) = \frac{\sum_{j=1}^{o} y(f_j, N_1) y(f_j, N_2)}{\sum_{j=1}^{o} y(f_j, N_1)^2 + \sum_{j=1}^{o} y(f_j, N_2)^2 - \sum_{j=1}^{o} y(f_j, N_1) y(f_j, N_2)}$$

where $y(f_j, N_1)$ and $y(f_j, N_2)$ are the weights of the semantic metadata component f_j in the two pages N_1 and N_2 respectively. The similarity score value ranges from 0 to 1. When the similarity score is 1 it means that the web pages have exactly same semantic metadata items and the similarity score 0 indicates that web pages have no similar semantic metadata items. The semantic similarity matrix M is generated using the similarity score value in such a way that each entry of M is the similarity score value between the two corresponding web pages.

3.2 Calculation of Socio Semantic Information

The terms which have high weight of the semantic metadata is collected and those terms are considered as most important terms in web pages. The semantic orientation between the most important terms and the terms in tweets are calculated by using Pointwise Mutual Information (PMI) (https://marcobonzanini.com/2015/05/17/mining-twitter-data-with-python-part-6-sentiment-analysis-basics/). The PMI is calculated by using the following equation:

$$PMI(t_i, t_j) = log\left(\frac{P(t_i \wedge t_j)}{P(t_i). P(t_j)}\right)$$

where $t_i$is the most important terms in web pages and $t_j$denotes the terms in tweets, $P(t_i)$ denotes the probability of observing the term $t_i$and $P(t_j)$ denotes the probability of observing the term t_j, $P(t_i \wedge t_j)$ denotes the probability of observing the terms t_i and t_j occurring together. For a given set of tweets T, the document frequency (DF) of the term is defined as the number of the tweets where the term occurs. The $P(t)$ and $P(t_i \wedge t_j)$ is calculated as follows:

$$P(t) = \frac{DF(t)}{|T|}$$

$$P(t_i \wedge t_j) = \frac{DF(t_i \wedge t_j)}{|T|}$$

The PMI score is the twitter score and it is added with the semantic similarity matrix. It provides socio semantic information for the web page recommendation. The socio semantically enriched matrix is given as follows:

$$w_{p_i p_j} = D\left(T_{p_{i,p_j}}\right) + \begin{cases} (1-\alpha) \times M_{p_i p_j} \times PMI(t_i, t_j), M_{p_i p_j} & > 0, \\ 0 \quad , \qquad\qquad M_{p_i p_j} & = 0, \end{cases}$$

where, $D\left(T_{p_{i,p_j}}\right)$ represents the quickest change detection of transition probability matrix determined by SSVLCPMC model, $M_{p_i p_j}$ represents the similarity score matrix, α denotes the semantic coefficient factor and $PMI(t_i, t_j)$ denotes the twitter score calculated between the most important terms in web pages and the terms in the tweets. Based on the above weight matrix the next link choice is created and the recommendations are generated.

4 Results and Discussion

In this section the performance of the proposed and existing methods for web page recommendation are evaluated in terms of prediction accuracy, precision, coverage, F1 measure and R measure. For the experimental purpose, DBpedia dataset from semantic web dog food web site is used.

4.1 Prediction Accuracy

Prediction accuracy is the measure of correctly recommended pages in all instances. It can be calculated by

$$Accuracy = \frac{(True\ positive + True\ negative)}{(True\ positive + True\ negative + False\ positive + False\ negative)}$$

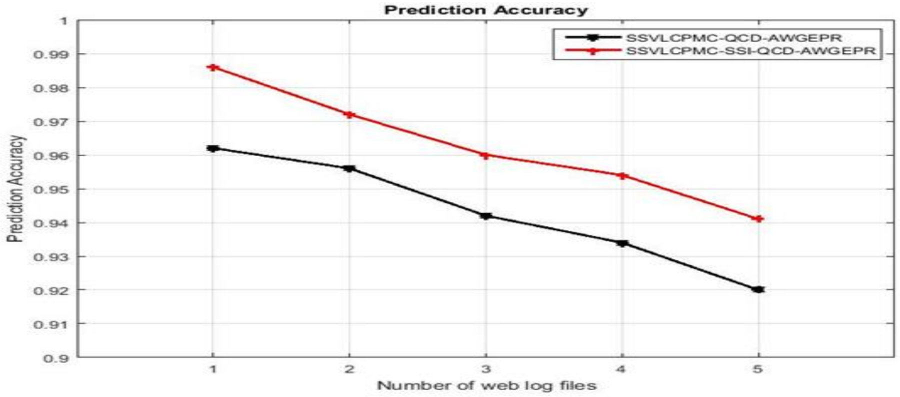

Fig. 1. Comparison of Prediction Accuracy

Fig.1, shows the comparison of accuracy between existing Support vector machine based Semantic Variable Length Confidence Pruned Markov Chain –Quickest Change Detection- Adaptive Weighted Gradient Estimation Page Rank (SSVLCPMC-QCD-AWGEPR) and proposed SSVLCPMC- Socio Semantic Information- QCD-AWGEPR (SSVLCPMC-SSI-QCD-AWGEPR) based web page recommendation methods. X axis denotes the number of web log files and Y axis denotes the prediction accuracy. From the Fig.1, it is proved that the proposed SSVLCPMC-SSI-QCD-AWGEPR shows high prediction accuracy than the existing SSVLCPMC-QCD-AWGEPR based web page recommendation method.

4.2 Precision

Precision is a number of relevant Web pages retrieved divided by the total number of Web pages in recommendations set. Thus precision of Rec with respect to t is given by,

$$Precision\ (Rec, t) = \frac{|Rec\ \cap (t-w)|}{|Rec|}$$

where, t denotes the user sessions from the test set, w denotes the window size, Rec denotes the recommendation set and $|Rec\ \cap (t-w)|$ denotes the number of common web pages in both recommendation set and evaluation set.

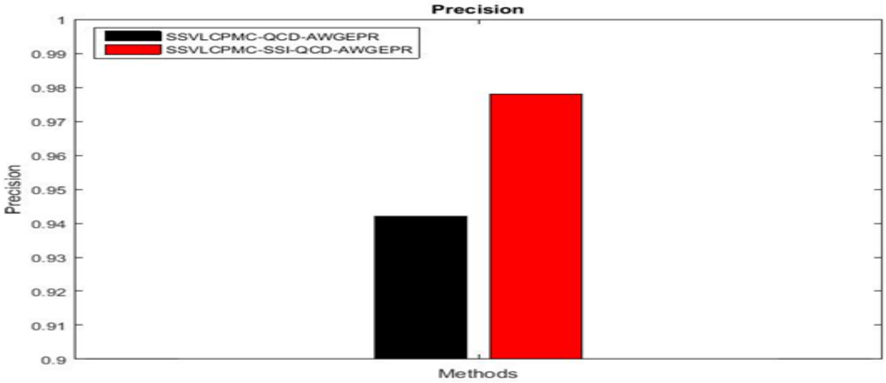

Fig. 2. Comparison of Precision

Fig.2 shows the comparison of precision between between existing SSVLCPMC-QCD-AWGEPR and proposed SSVLCPMC-SSI-QCD-AWGEPR based web page recommendation methods. X axis denotes the methods and Y axis denotes the precision. From the Fig.2, it is proved that the proposed SSVLCPMC-SSI-QCD-AWGEPR shows high precision than the existing SSVLCPMC-QCD-AWGEPR based web page recommendation method.

4.3 Coverage

Coverage is the ratio between the number of relevant Web pages retrieved and the total number of Web pages that actually belongs to the test user session. Coverage of Rec with respect to t is given by,

$$coverage\ (Rec, t) = \frac{|Rec\ \cap (t-w)|}{|t-w|}$$

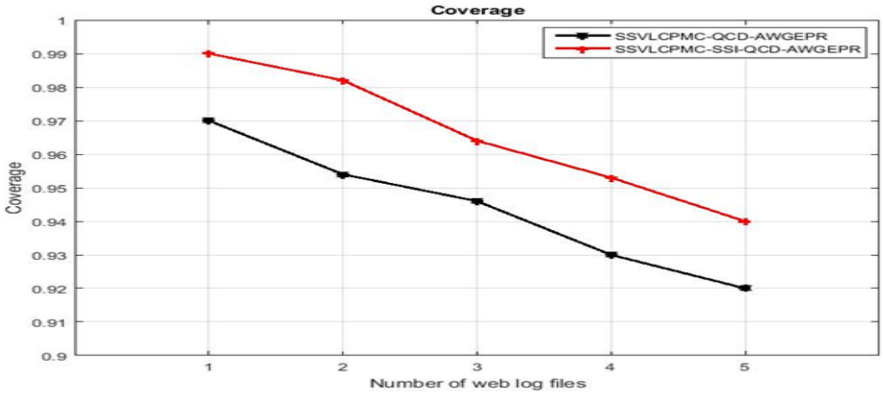

Fig. 3. Comparison of Coverage

Fig.3, shows the comparison of coverage between existing SSVLCPMC-QCD-AWGEPR and proposed SSVLCPMC-SSI-QCD-AWGEPR based web page recommendation methods. X axis denotes the number of web log files and Y axis denotes the coverage. From the Fig. 3, it is proved that the proposed SSVLCPMC-SSI-QCD-AWGEPR shows high coverage than the existing SSVLCPMC-QCD-AWGEPR based web page recommendation method.

4.4 F1 Measure

F1 measure is used to achieve high precision and high coverage. F1 measure is given by,

$$F1\ (Rec,t) = \frac{2 \times precision\ (Rec,t) \times coverage\ (Rec,t)}{precision\ (Rec,t) + coverage\ (Rec,t)}$$

F1 measure achieves its maximum value when both precision and coverage are maximized.

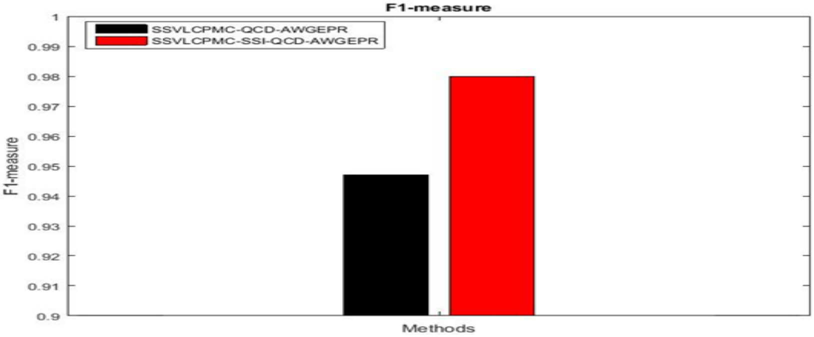

Fig.4 . Comparison of F1-measure

Fig. 4, shows the comparison of F1-measure between existing SSVLCPMC-QCD-AWGEPR and proposed SSVLCPMC-SSI-QCD-AWGEPR based web page recommendation methods. X axis denotes the methods and Y axis denotes the F1-measure. From the Fig.4, it is proved that the proposed SSVLCPMC-SSI-QCD-AWGEPR shows high F1-measure than the existing SSVLCPMC-QCD-AWGEPR based web page recommendation method.

4.5 R measure

R measure is evaluated by dividing the coverage by the size of the recommendation set and it is given by,

$$R\ (Rec,t) = \frac{coverage\ (Rec,t)}{|\ Rec\ |}$$

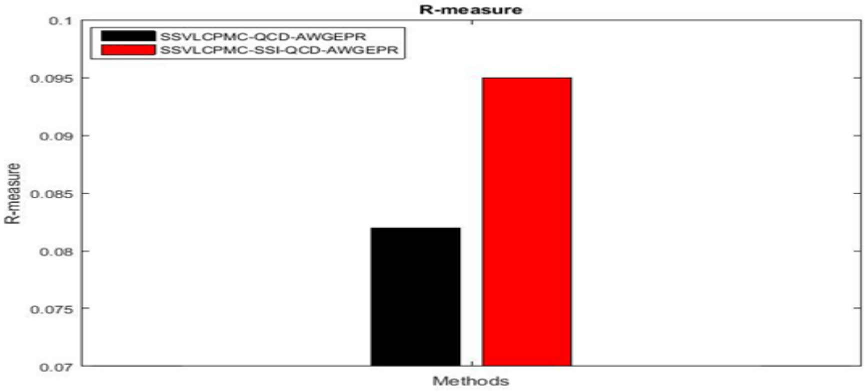

Fig. 5 Comparison of R-measure

Fig. 5, shows the comparison of R-measure between existing SSVLCPMC-QCD-AWGEPR and proposed SSVLCPMC-SSI-QCD-AWGEPR based web page recommendation methods. X axis denotes the methods and Y axis denotes the F1-measure. From the Fig. 5, it is proved that the proposed SSVLCPMC-SSI-QCD-AWGEPR shows high R-measure than the existing SSVLCPMC-QCD-AWGEPR based web page recommendation method.

5. Conclusion

In this paper, web page recommendation system is further improved by considering the socio semantic information. The semantic similarity between collected web pages is calculated by using TF-IDF. It produces the composite weight composite weight for each semantic metadata in the web page. The most important terms in the web pages are obtained based on the composite weight. Then the semantic orientation between the important terms and the terms in tweets are calculated using PMI which returns a twitter score. This score is included with the semantic similarity matrix which improves web page recommendation accuracy. The experimental results shows that the proposed web page recommendation method achieves high prediction accuracy, precision, coverage, F1-measure and R-measure than the existing web page recommendation method.

References

Bhavsar, M., & Chavan, M. P. (2014). Web page recommendation using web mining. *Int. Journal of Engineering Research and Applications*, *4*(7), 201-206.

Waykule, V., & Gupta, S. S. (2014). Review of Web Recommendation System and Its Techniques: Future Road Map. *International Journal of Computer Science and Information Technologies*, *5*(1), 547-551.

Shirgave, S., Kulkarni, P., & Borges, J. (2014). Semantically Enriched Variable Length Markov Chain Model for Analysis of User Web Navigation Sessions. *International Journal of Information Technology & Decision Making*, *13*(04), 721-753.

Nguyen, T. T. S., Lu, H. Y., & Lu, J. (2014). Web-page recommendation based on web usage and domain knowledge. *IEEE Transactions on Knowledge and Data Engineering*, *26*(10), 2574-2587.

Wen, H., Fang, L., & Guan, L. (2012). A hybrid approach for personalized recommendation of news on the Web. *Expert Systems with Applications*, *39*(5), 5806-5814.

Jalali, M., Mustapha, N., Sulaiman, M. N., & Mamat, A. (2010). WebPUM: A Web-based recommendation system to predict user future movements. *Expert Systems with Applications*, *37*(9), 6201-6212.

Abrishami, S., Naghibzadeh, M., & Jalali, M. (2012, December). Web page recommendation based on semantic web usage mining. In *International Conference on Social Informatics* (pp. 393-405). Springer, Berlin, Heidelberg.

Rizvi, N. T. S. H., & Keole, R. R. (2015). Web page recommendation in information retrieval using domain knowledge and web usage mining. *International Journal of Science, Engineering and Technology Research*, *4*(5), 1531-1535.

Ramesh, C., Rao, K. V., & Govardhan, A. (2011). A semantically enriched web usage based Recommendation model. *International Journal of Computer Science & Information Technology (IJCSIT), 3(5),* 193-202.

Rizvi, N. S., & Keole, R. R. (2015). Use Of Ontology And Web Usage Mining For Web-Page Recommendation. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* 4(4), 1095-1099.

Nguyen, T. T. S., Lu, H. Y., & Lu, J. (2010, November). Ontology-style web usage model for semantic web applications. In *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on* (pp. 784-789). IEEE.

https://marcobonzanini.com/2015/05/17/mining-twitter-data-with-python-part-6-sentiment-analysis-basics/.