WWJMRD 2018; 4(4): 70-78
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
UGC Approved Journal
Impact Factor MJIF: 4.25
E-ISSN: 2454-6615

D.Dhayalan
Research scholar, Department of Computer Science, Dravidian University kuppam, A.P, India

Dr.C.Chandrasekar
Professor, Department of Computer Science, Periyar University, Salem, Tamil Nadu, India

Correspondence:
D.Dhayalan
Research scholar, Department of Computer Science, Dravidian University kuppam, A.P, India

Mining Semantic Relation Using Polynomial Kernel with Support Vector Machine from Textual Web Content

D.Dhayalan, Dr.C.Chandrasekar

Abstract
Traditional text mining techniques for semantic association transform free text into association semantic and Points of Interests (POI) for knowledge discovery. However, it does not consider time and precision while performing linear inseparability from textual web content. Effective relation association semantic technology using machine learning analysis not only can reduce time for knowledge discovery from textual web content, but also can improve the searching accuracy of the related information system. How to realize the semantic relation mining by the machine learning analysis is an important research topic. In this paper, a three-step procedure to mine associations of semantic relations for textual web document content, called, Polynomial Kernelized Maximum Entropy and Support Vector Machine (PKME-SVM) framework is presented. First, Associative Polynomial Kernel-based Maximum Entropy representing semantic relations are extracted from raw text web contents using Polynomial Kernel. The semantic relation extraction process also creates a sentence grammar tree in the form of reduced sentence. Then, Probabilistic Term (i.e. word) Taxonomy (PTT) Framework is applied to discover the probabilistic term on corresponding Web Content Domain. Finally, for pruning the semantic relation from textual web content, the Generalized Association Support Vector Mining algorithm adopts the notion of dual characteristic function for systematic overgeneralization reduction. The objective of PKME-SVM is to obtain accurate result from textual web content using semantic relation mining operations and satisfy the web user specific needs through polynomial kernel. The efficacy of our framework is demonstrated through empirical experiments conducted on Freebase Data Dump. Experiment is conducted on the factors such as personalized information search retrieval rate, computation time and precision ratio.

Keywords: Text mining, Points of Interests, Associative Polynomial Kernel, Maximum Entropy, Probabilistic Term Taxonomy, Web Content Domain

1. Introduction

Association link network (ALN) refers to semantic link network constructed by mining the association relations between Web resources for efficient supporting of applications related to web like, Web semantic association search, Web knowledge discovery and so on. Hierarchy Cutting model [1] of association semantic was presented with the core idea of realizing the layer design to not only ensure theoretical support, but also improve the searching efficiency. This was achieved by initially discovering the possible distribution law, followed by division of association semantic into three layers and finally analyzing the domain topic was presented.

Despite high accuracy being achieved, the computation time for analysing domain topic on the web remained unaddressed. To address this solution, Associative Sentence Polynomial Kernel-based Maximum Entropy is applied to the raw text web contents to that the time taken to pre-process is said to be reduced and therefore the overall computation time for analyzing domain topic on the web.

A novel Geosocial Semantic Personalized Location Information System (G-SPLIS) [2] to provide Location-Based Social Networking Services (LBSNSs) was presented to ensure qualitative contextualized information to their users. The G-SPLIS using LBSNS

Formalized human daily preferences. Point Of Interest owners' group targeted policies

were also ensured with the potentiality to update user-defined logical rules. Owing to the fact that these rules to be distributed between several systems, such user-defined policies were found to be straightforwardly reused and expanded, ensuring personalization problem despite median and average time required for completing each task was said to be reduced, focus want not made on the rate of precision. In this work, to address this issue, Probabilistic Term Taxonomy Framework based on approximated likelihood ratio is presented with the objective of improving the rate of precision and recall in addition to the average computation time for analyzing domain topic on the web. A framework for semantic social media analysis was investigated in [4]. Latent semantic analysis was applied in [5] for short text mining. A systematic literature review by text and association rule mining was presented in [6].

From the above papers, it can be seen that a major assumption is that source data were designed on the basis of user defined policies. This, however, would reduce the usability of these algorithms, as it requires the users to have a well understanding of the source data or domain topic especially on the web. In addition, better pre-processing of web content is required, which is significantly difficult to be obtained.

In this paper, we propose a novel algorithm, called PK-Maximum Entropy algorithm (Polynomial Kernel). The PK-Maximum Entropy algorithm differs itself from the traditional methods by utilizing polynomial kernels to strengthen the text document analysis for raw text web contents without the need of specifying the theoretical support. Specifically, using the machine learning analysis techniques, the PK-Maximum Entropy algorithm first extracts the sentence grammar as the seed feature set between the raw text web contents and the target data. This is done by employing the Associative Sentence Polynomial Kernel-based Maximum Entropy (ASPK-ME) model to get the most semantic relations from textual web content, obtaining reduced sentence.

The Probabilistic Term Taxonomy algorithm then performs lexical taxonomy relation and taxonomy construction for the web content domain and reduced sentence. With such a term/ semantic relation extraction in hand, the Generalized Association Support Vector Mining algorithm further mines semantic association for web topic domain analysis. Finally, the classification has been done through minimizing the optimal term-document matrix between the instance and feature representation using dual characteristic function. Our major contributions are as follows:

- We propose the Generalized Association Support Vector Mining algorithm, a machine learning algorithm of effective semantic relation mining. The Generalized Association Support Vector Mining is superior to other algorithms in terms of automatically identifying the probabilistic term from the rich raw text web contents with no requirements of the priori probability distribution and it integrates the polynomial kernel support vector machine into mining accurate semantic relation which facilitates an effective learning.
- We conduct extensive experimental evaluations and the experimental results indicate that our proposed framework is effective.
- We find that the Generalized Association Support Vector Mining algorithm may be applied to semantic relation mining from textual web content.

The reminder of the paper is organized as follows. In section 2, related work is reviewed. Then in section 3, the details of the Polynomial Kernelized Maximum Entropy and Support Vector Machine framework is presented. Followed by which in section 4, experimental evaluations are investigated. In section 5 discussions for various parameters are provided with the aid of graph. Finally, conclusion is included in section 6.

2. Related works

Information outburst highlights the requirements for machines to better comprehend the natural language texts. In [7], semantic knowledge was analyzed based on short text using co-occurrence network. With the co-occurrence data, online customer reviews are gaining great range of popularity. To address this issue, in [8], a text processing framework was designed based on unsupervised method resulting in better accuracy of customer reviews being retrieved. Yet another work with the objective of improving accuracy for social tag analysis was presented using joint latent dirichlet allocation was presented in [9]. A unified framework was designed in [10] using Bayesian-based framework.

Today, the epidemic development of social networking data, integrated with developments in the area of semantic techniques, is giving rise to several fascinating research topics. In [11], several semantics and knowledge engineering with social networks were investigated. Yet another conceptual model for web personalization through semantic annotation system was presented in [12]. However, with the explosive data growth, mining semantic relations remains a major task as far as time is concerned. To address this issue, cognitive inspired for knowledge discovery was presented in [13]. However, with the inclusion of multiple domains, mining yet remains a complicated issue. In [14], Resource Description Framework graphs to avoid server overload with minimal delay between sequential requests was presented.

With the ever growing increase in biomedical publications, managing literature in an efficient manner to support hypothesis generation and discovery has become an overwhelming task. Text mining has been proposed to address this issue. The area of semantic analysis for different applications has been increased in the recent years.

A rule-based compositional approach using lexical and syntactic information was presented in [15], resulting in the improvement of semantic predications being extracted. In [16], Natural Language Processing (NLP) techniques were used with the objective of extracting the semantic and syntactic features with respect to chemical exposure text, facilitating higher information retrieval rate.

In today's scenario, users spend maximum time doing collective tasks at anywhere and anytime. Especially, professionals and researchers want to collectively work with each other with the aid of certain advanced technologies to share knowledge with the objective of improving the business processes. In [17], semantic web technologies were investigated using Service Oriented Architecture (SOA) based solution collaborate independently regarding technical details and specific

context tasks in a decentralized manner. However, the SOA based architecture was limited towards business process.

To perform geological knowledge extraction and identifying the semantic association between geological data, Natural Language Processing (NLP) and Data Mining (DM) algorithms were applied in [18]. With this large number of geological documents were extracted in a dynamic manner. Despite, accuracy attained during mining using geological documents, issues related to unstructured documents were not analyzed.

To address this issue, a semantic annotation strategy for unstructured documents was presented. Here, ontologies [19] were used to identify the context of the entities specified in the query which therefore resulted in the improvement of rate of concept similarity. A systematic mapping related to semantics-concerned text mining was presented in [20] that paved the way for guiding the researchers working with semantics-concerned text mining.

Based on the aforementioned methods and mechanisms, in this paper, an efficient framework called, Polynomial Kernelized Maximum Entropy and Support Vector Machine (PKME-SVM) is presented to perform semantic relation mining from textual web content. The elaborate description of this is provided in the forthcoming sections.

3. Polynomial Kernelized Maximum Entropy and Support Vector Machine framework

Semantic relationships represent the associations that prevail between the meanings of words, between the meanings of phrases, or between the meanings of sentences. Following is a description of such relationships. Pairs of words that are synonymous are supposed to part all their semantic features or characteristics. Nevertheless, no two words possess exactly similar definition in all the contexts, where they occur.

In this paper, a novel framework called, Polynomial Kernelized Maximum Entropy and Support Vector Machine is used to improve the accuracy and time taken for association mining of semantic relations from textual web content by modifying the existing hierarchy cutting and context-aware web mapping. The proposed framework provides various advantages such as personalized information search retrieval rate or accuracy, computation time and precision ratio from textual web content. Figure 1 given below shows the block diagram of Polynomial Kernelized Maximum Entropy and Support Vector Machine framework.

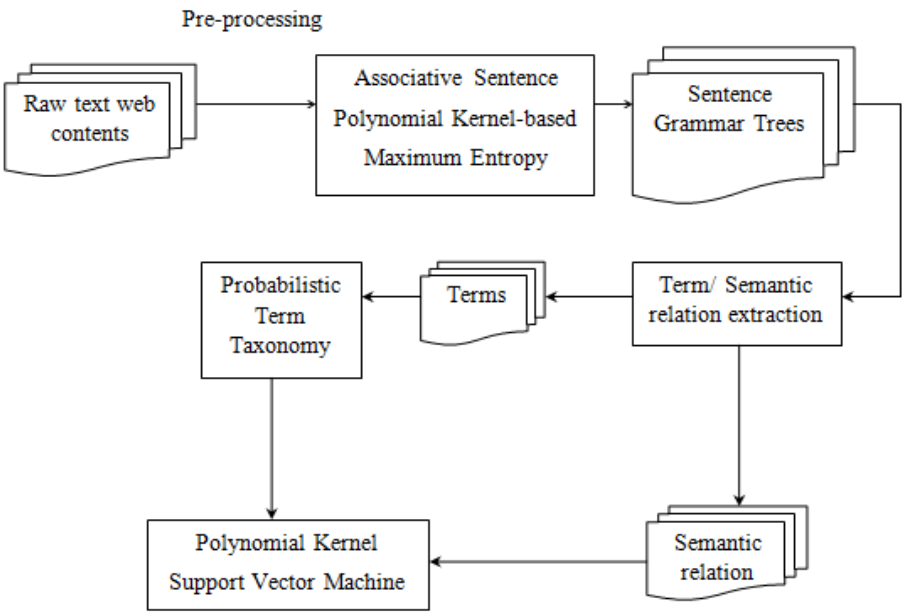

Fig.1: Block diagram of Polynomial Kernelized Maximum Entropy and Support Vector Machine framework

As shown in the figure, raw text web contents extracted from Freebase Data Dump is provided as input. Based on the preprocessing results obtained through Associative Sentence Polynomial Kernel-based Maximum Entropy, a set of predefined rules adopted from [4] is applied for efficient extraction of semantic relations from sentence grammar trees. As shown in the figure, during the semantic relation extraction stage, raw text web contents applied as input to be pre-processed using Polynomial Kernel. This in turn produces text documents with Polynomial Kernel. To the text document output with polynomial kernel Maximum Entropy model.

After pre-processing, each parsed text document includes a set of sentence grammar trees. According to the sentence grammar trees, terms and semantic relations are extracted by applying Probabilistic Term Taxonomy Framework and Polynomial Kernel Support Vector Machine model respectively. This proposed PKME-SVM has been implemented using JAVA platform and proved that the proposed algorithm provides better performance when it is compared with the other existing algorithms. The elaborate description of Polynomial Kernelized Maximum Entropy and Support Vector Machine framework is given below.

3.1 Associative Polynomial Kernel-based Maximum Entropy model

The first step in the association mining of semantic relations from textual web content is pre-processing. In this work, an Associative Sentence Polynomial Kernel-based Maximum Entropy (ASPK-ME) model is generated for sentence grouping. Figure 2 given below shows the flow diagram of Associative Polynomial Kernel-based Maximum Entropy (ASPK-ME) model.

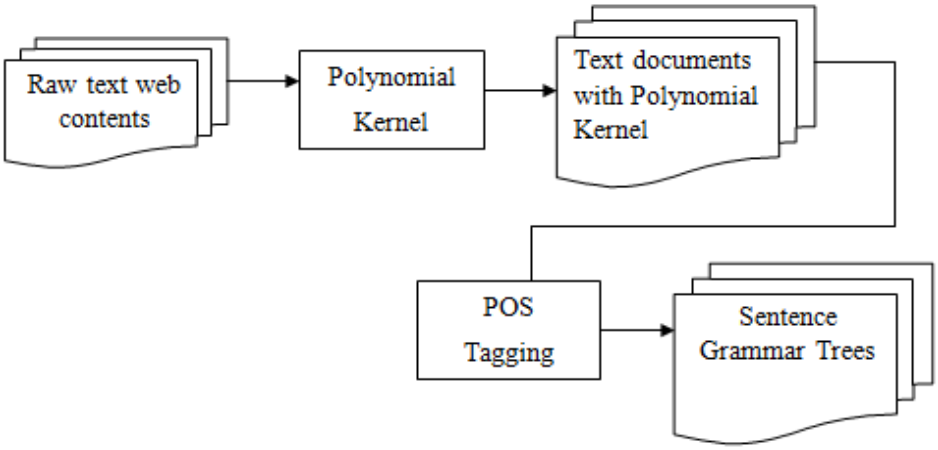

Fig.2: Flow diagram of pre-processing using Associative Polynomial Kernel-based Maximum Entropy

In the fundamental vector space representation, association sentences are denoted by a matrix 'M', whose columns are indexed by the sentences and rows are indexed by the words. A sentence 'S' with words tagged for part of speech is denoted by a row vector i.e, '$S = < (w_1, t_1), (w_2, t_2), \dots, (w_n, t_n) >$'. For POS tagging, the proposed ASPK-ME model has applied based on Maximum Entropy [3] model. It is mathematically written as given below.

$$\emptyset(S) = [tf(t_1, S), tf(t_2, S), \dots, tf(t_n, S)] \quad (1)$$

From the above equation (1), '$tf(t_i, S)$' represents the frequency of term 'i' occurred in sentence 'S'.The analogous kernel 'K' is then given by the central product between the feature vectors, as given in equation (2) and (3).

$$K = M'M \quad (2)$$

$$k(S_1, S_2) = < \emptyset(S_1), \emptyset\,(S_2) > = \sum_{i=1}^{n} tf(t_i, S_1), tf(t_i, S_2) \quad (3)$$

From the above two equations, the pre-processed text is extracted with minimum time for degree 'd' polynomials and is mathematically written as given below.

$$k(S_1, S_2\,) = (\emptyset(S_1\,), \emptyset\,(S_2\,))\hat{}d \quad (4)$$

From the above equation (4), the similarity of words between two sentences 'S_1' and 'S_2' denotes their similarity value. Given the sentence 'S', the reduced sentence 'S'' represents the subsequence of original sentence 'S', that is obtained by removing punctuations based on Maximum Entropy. The pseudo code representation of PK-Maximum Entropy is as given below (algorithm 1).

Input: Web Content Domain 'WCD', Sentences '$S_1, S_2, \dots, S_n$', User Query '$UQ = 〖UQ〗_1, 〖UQ〗_2, \dots, 〖UQ〗_n$'
Output: reduced sentence 'S^''
1. **Begin** 2. **For** each Sentences 'S_1' and Web Content Domain 'WCD' with User Query 'UQ' 3. Measure the frequency of term 'i' occurred in sentence 'S' using equation (1) 4. Obtain the kernel using the equation (2) 5. Measure the kernel for degree 'd' polynomials using equation (4) 6. End for 7. End

Algorithm 1 PK-Maximum Entropy algorithm

As given above the, the raw text web contents are pre-processed by applying the PK-Maximum Entropy algorithm. This is performed by applying the Polynomial Kernel and the Maximum Entropy model. For each sentence, the PK-Maximum Entropy algorithm eliminates the ambiguity of pronouns in text by applying the Polynomial Kernel. Followed by which, Maximum Entropy is applied to the resultant text to obtain the parts of speech that forms the most important part of semantic relations from textual web content.

3.2 Probabilistic Term Taxonomy Framework

The second step in the association mining of semantic relations from textual web content is the construction of Probabilistic Term (i.e. word) Taxonomy (PTT) Framework that performs the task of taxonomy relation identification and taxonomy construction. The objective remains in constructing a term (or word) taxonomy so as to group similar terms or words into relevant groups. Based on such relevant groups, semantic relations that correspond to similar terms or words are then hypothesized for obtaining statistically important sentences during the knowledge mining stage.

In PTT framework, given an ordered pair of two terms 'T_1' and 'T_2', lexical taxonomy relation identification is mathematically formulated as given below.

$$Score_{LTR} = \frac{Log\,(|C_{WCD}(T_1, T_2)|)}{1 + Log\,(|C_{WCD}(T_2, T_1)|)} \quad (5)$$

For each term or word 'W' in the extracted reduced sentence a search for a term taxonomy is first made. From the extracted reduced sentence 'S'', '$S'C$' representing the set of singleton word components appearing in '$S'T$' of a Web Content Domain 'WCD', a Probabilistic Term Taxonomy is mathematically obtained as given below.

$$A(S'(T)) = \frac{Prob\,(T)}{\sum_{i=1}^{n} Prob\,(T_i)} \quad (6)$$

From the above equation (6), '$A(S')$' corresponds to the

approximated likelihood of term '(T)' in Web Content Domain'WCD'.If '$S' = \{data\ base\ management, management\ method\}$', then '$S'(T) = \{data, base, management, method\}$' and '$A(S'(T)) = 2/4$'. It is diagrammatically represented as in figure 3.

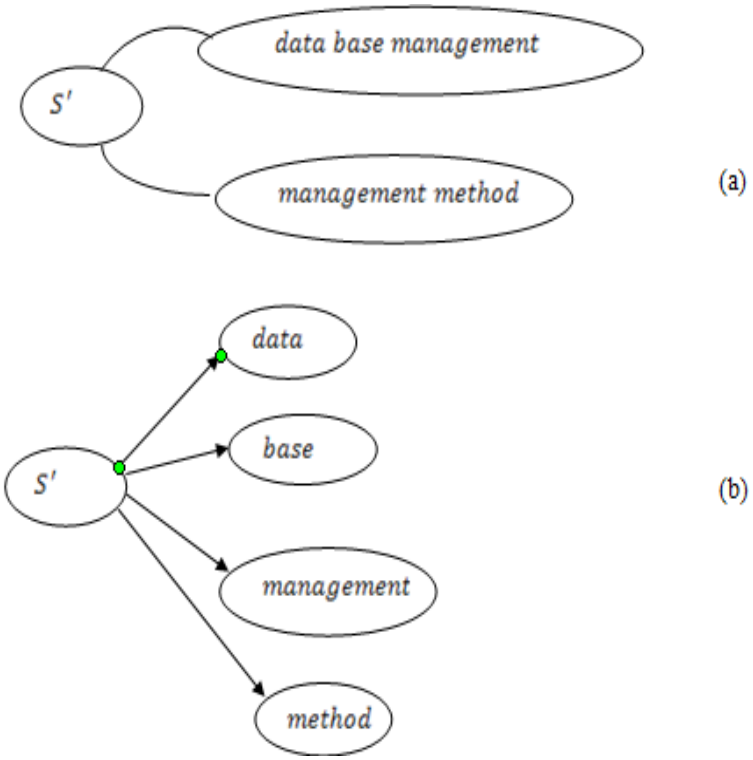

Fig.3: (a) Conversion of Probabilistic Term Taxonomy into (b) conceptual graph

As illustrated in the figure, the conversion of Probabilistic Term Taxonomy into conceptual graph results in the improvement of precision and recall rate with the precise term taxonomy generated through probabilistic model. The pseudo code representation of Probabilistic Term Taxonomy is given below (algorithm 2).

Input: Web Content Domain 'WCD', reduced sentence 'S'', User Query '$UQ = UQ_1, UQ_2, \ldots, UQ_n$'
Output: Probabilistic Term $A(S'(T))$
1. **Begin** 2. **For** Web Content Domain 'WCD' and reduced sentence 'S'' with User Query 'UQ' 3. Identify lexical taxonomy relation using equation (5) 4. Measure Probabilistic Term Taxonomy using equation (6) 5. **End for** 6. **End**

Algorithm 2 Probabilistic Term Taxonomy algorithm

As given above, the Probabilistic Term Taxonomy algorithm includes two parts. The first part identifies the lexical taxonomy relation in a dynamic manner with respect to Web Content Domain for the corresponding reduced sentence. Followed by this, the second part measures the Probabilistic Term Taxonomy based on the approximated likelihood ratio. The identification of lexical taxonomy relation and taxonomy construction therefore results in the improvement of precision and recall.

3.3 Polynomial Kernel Support Vector Machine (PK-SVM) model

Finally, a novel web textual association mining, called Generalized Association Support Vector Mining algorithm is applied to identify the latent relation association of semantic relations on probabilistic term. According this model, the PK-SVM classifies a given unknown Probabilistic Term by the following classification decision formula.

$$(p_i, q_i), \ldots . (p_n, q_n), p \in A(S'(T)), q \in \{-1, +1\} \quad (7)$$

From the above equation (7), '$(p_i, q_i), \ldots . (p_n, q_n)$' Corresponds to the training samples with 'n' representing the total number of samples, whereas 'q' represents the resultant category '$\{-1, +1\}$' respectively. The classification decision formula used in the proposed work is as given below.

$$(w.p_i) + b > 0\ if\ q = +1 \quad (8)$$

$$(w.p_i) + b < 0\ if\ q = -1 \quad (9)$$

The SVM includes different kernel functions to solve different classification problems. In this work, Polynomial Kernel is selected to solve the problem of linear inseparability. Let us assume a term document matrix 'TDM', which is a '$n * d$' matrix. Here 'n' corresponds to the number of words and 'd' corresponds to the number of documents. Each element '$TDM[n, d]$' refers to the number of existences of word 'n' in document 'd'. Then, the term document matrix 'TDM' of SVD is defined as given below.

$$TDM = PSQ^A \quad (10)$$

$$TDM_k = P_k S_k Q_k^A \quad (11)$$

From the above equation (10) and (11), 'S' corresponds to the singular values of term document matrix 'TDM', 'P_k' corresponds to the web content vector and 'Q_k' corresponds to the term vector. Since, the objective of the work remains in mining the common semantic relation from textual web content, the Polynomial Kernel Support Vector Machine (PK-SVM) model only processes textual web content vector and top 'k' singular value matrix is selected, as it provides the latent semantic relationship. This is mathematically formulated as given below.

$$Res\ Value = P_k * S_k \quad (12)$$

From the above equation (12), the semantic feature vector of each textual web contents are retrieved. The dual characteristic function 'A_α^m' is then mathematically defined as given below.

$$A_\alpha^m\ (TDM_i) = \begin{cases} 1, SF_\alpha\ (TDM_i)\ \in\ \beta_m \\ -1, SF_\alpha\ (TDM_i)\ \notin \beta_m \end{cases} \quad (13)$$

From the above equation (13), 'SF_α' represents the semantic features with 'TDM_i' representing the term document matrix for web content domain, with 'α' representing the index model of SVM for 'β_m' type of categories. The pseudo code representation of Generalized Association Support Vector Mining algorithm is given below (algorithm 3).

Input: Web Content Domain 'WCD', reduced sentence 'S'', Probabilistic Term $A(S'(T))$, User Query '$UQ = UQ_1, UQ_2, \ldots, UQ_n$'
Output: accurate semantic relation from textual web content
1. **Begin** 2. **For** each Web Content Domain 'WCD' with Probabilistic Term $A(S'(T))$ with User Query 'UQ' 3. Obtain term document matrix using equation (10) 4. Obtain latent semantic relationship using equation (12) 5. Evaluate the dual characteristic function using equation (13) 6. **End for** 7. **End**

Algorithm 3 Generalized Association Support Vector Mining algorithm

As shown in the above Generalized Association Support Vector Mining algorithm, for each Web Content Domain 'WCD' with Probabilistic Term '$A(S'(T))$', first term document matrix is evaluated with which the problem of linear inseparability is solved where, the high term-document matrix dimensions are reduced to optimal term-document matrix. Next, latent semantic relationship is obtained by using the web content vector. Finally, by applying the dual characteristic function, substantially reduces the term redundancy and perform much better than the conventional semantic association for analyzing web domain topic.

4. Experimental setup

In this section, we conducted the experiment in JAVA platform database to evaluate the performance of the proposed Polynomial Kernelized Maximum Entropy and Support Vector Machine (PKME-SVM) framework for mining semantic relation from web content. Experiments are conducted using the Freebase Data Dump due to the presence in several domains and hence are analyzed to identify the efficiency level so that the performance is said to be examined from web content. The proposed framework is evaluated with the aid of Freebase Data Dump with the updated user profiles.

A data dump includes essential information for mining the facts related to each subject in Freebase. Freebase is an open source database that covers millions of theme in hundreds of group and the PKME-SVM framework fetches the accurate result to the users. Drawing from large open datasets like Wikipedia, MusicBrainz, and the SEC archives simultaneously contain prearranged information on several popular topics, including movies, music, people and locations. The information includes the historic events, European railway stations and chemical properties of common food ingredients. Present research work is compared against the existing Hierarchy Cutting model [1] and Geosocial Semantic Personalized Location Information System (G-SPLIS).

5. Discussion

Three widely used metrics are used in the experiments. The three metrics included are, personalized information search retrieval rate, computation time and precision.

5.1 Impact of accuracy

The personalized information search retrieval rate refers to the accuracy of user request retrieval which is given as follows. It is measured in terms of percentage (%). The personalized information search retrieval rate refers to the accuracy of user request retrieval from textual web content using the Freebase Data Dump which is given as follows. It is measured in terms of percentage (%).

$$A = \sum_{i=1}^{n} \frac{UR_i}{n} \quad (14)$$

From the above equation (14), 'UR_i' is the user request made and n is the number of total samples made in the testing set and with the ratio of user request and total samples, accuracy is said to be arrived at. As shown in the figure, the performance of Polynomial Kernelized Maximum Entropy and Support Vector Machine (PKME-SVM) framework from textual web content for mining semantic relation is compared with the existing Hierarchy Cutting model [1] and Geosocial Semantic Personalized Location Information System (G-SPLIS). The performance is evaluated according to the following metrics.

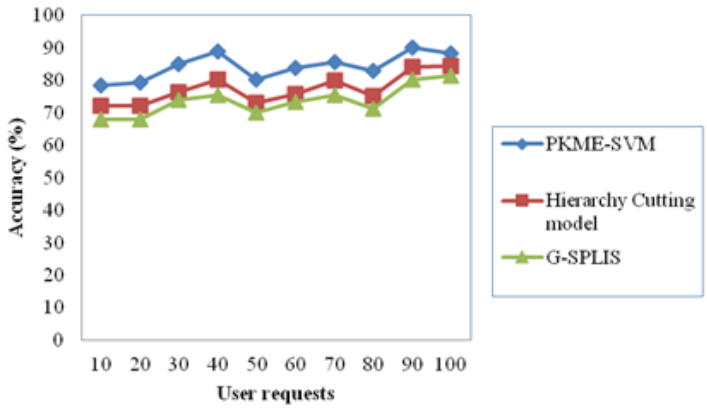

Fig.4: Accuracy of user request retrieval from textual web content analysis

Figure 4 shows the result of accuracy of user request retrieval from textual web content analysis versus the varying number of user requests based on the user requests made regarding different movies provided in Freebase Data Dump. To better perceive the efficacy of the proposed PKME-SVM framework, substantial experimental results are illustrated in Figure 4 and compared against the existing Hierarchy Cutting [1] model and G-SPLIS [2] respectively. Results are presented for different number of user requests that cover millions of movies related to different theme in hundreds of group.

The accuracy rate for semantic relation mining from textual web content user requests are performed at different time interval is shown above. Higher, the number of user requests being sent, more successful the method is and at the same time, user requests are not directly proportional to the accuracy rate.

As shown above, improvement with respect to accuracy is found for the user requests with 30 and 40 and again increase in accuracy with user request from 80 – 100. This is because with the including a set of sentence grammar trees, if properly separated into terms and relations, accuracy is said to be increasing rate. However, this is not found to be tree for all user requests. However, comparatively, proper set of sentence grammar trees was formed by applying the probabilistic term taxonomy algorithm. Therefore improvement found to be by applying the proposed PKME-SVM framework by 8% compared to Hierarchy Cutting [1] model. The process is repeated for 100 user requests for conducting experiments. In addition, by separating the probabilistic term into identification of taxonomy relation and taxonomy construction, the accuracy rate is said to be improved by 14% compared to G-SPLIS respectively.

5.2 Impact of computation time

The computation time is the time taken to perform text documents with polynomial kernel and sentence grammar trees with which the association semantic mining from text web content is said to be performed. The computation time therefore for obtaining sentence grammar is mathematically formulated as given below and is measured in terms of milliseconds.

$$C_{time} = Time\ (k(S_1, S_2)) \tag{15}$$

From the above equation (15), the computation time for extracting sentence grammar is obtained for 'd' degree polynomials based on the similarity of words between two sentences 'S_1' and 'S_2'. In order to reduce the computation time for knowledge discovery during linear inseparability from textual web content, user requests and user profile information for analyzing domain topic on web is considered according to different music extracted from Freebase Data Dump. In the experimental setup, the number of user requests ranges from 10 to 100 is illustrated in figure 5. The computation time using the framework PKME-SVM provides comparable values than the state-of-the-art methods.

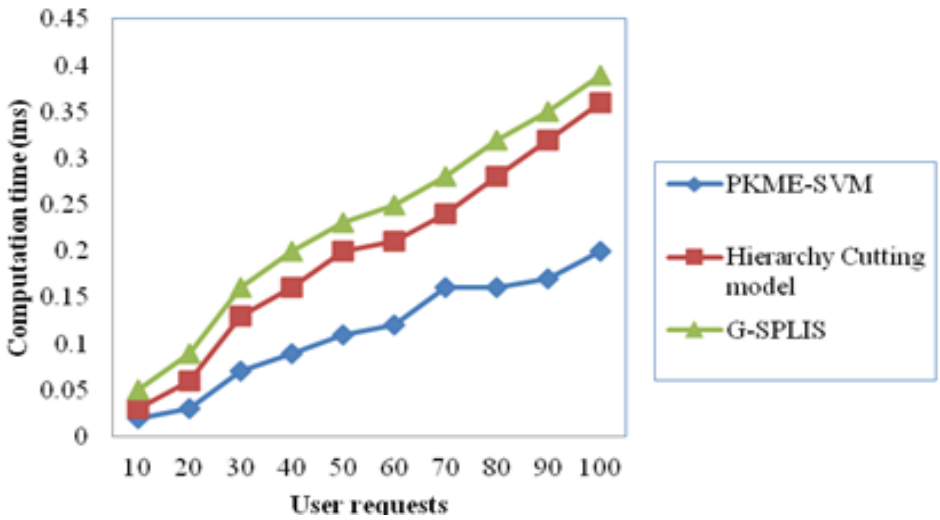

Fig.5: computation time of user request retrieval from textual web content

The targeting results for analyzing web content domain to measure the computation time using PKME-SVM framework is compared with two state-of-the-art methods Hierarchy Cutting model and G-SPLIS in figure 5 is presented for visual comparison based on the number of user requests being sent for extracting different music. Our framework PKME-SVM differs from the Hierarchy Cutting [1] model and G-SPLIS [2] in that we have incorporated association sentences that employ Associative Sentence Polynomial Kernel-based Maximum Entropy (ASPK-ME) model in semantic web mining. With the objective of reducing the computation time in PKME-SVM, different semantic relations from textual web content is based on the Maximum Entropy model.

For each sentence, the PK-Maximum Entropy algorithm removes the ambivalence of pronouns in text via Polynomial Kernel with maximum accurate result rate is compared and evaluated which helps in the minimization of computation time by 65% compared to Hierarchy Cutting model. Furthermore, with the effective application of PK-Maximum Entropy algorithm, reduced sentence is obtained through vector space representation by 25% compared to G-SPLIS.

5.3 Impact of precision

Precision refers to the closeness of two or more measurements to each other. For example, if the actual occurrences of a term in a document are 50, but the occurrences of term in a document is 35 then the measurement is not accurate. In this case, the measured value is not closed to the actual value. Using the example above, if you measure the occurrences of a word five times, and get 10 each time, then the measurement is very precise. Hence, precision is independent of accuracy. Hence, precision is the ratio of the mining of relevant results extracted to the total number of irrelevant and relevant results extracted for semantic relation mining from textual web content. It is usually expressed as a percentage. It is given as

$$P = \frac{UR_R}{UR_R + UR_I} \tag{16}$$

From the above equation (16), the precision ratio 'P' is arrived at for raw text web contents, based on the Sentences 'S_1' and Web Content Domain 'WCD' and user request provided as input. As discussed above, precision is independent of accuracy and is measured in terms of percentage (%). Figure 6 given below shows the rate of precision arrived at for different user requests for PKME-SVM framework, Hierarchy Cutting [1] model and G-SPLIS [2] versus 100 different user requests made regarding the location of the origin of music and movies. The precision ratio returned over PKME-SVM framework increases gradually though not linear for differing user requests for semantic web mining.

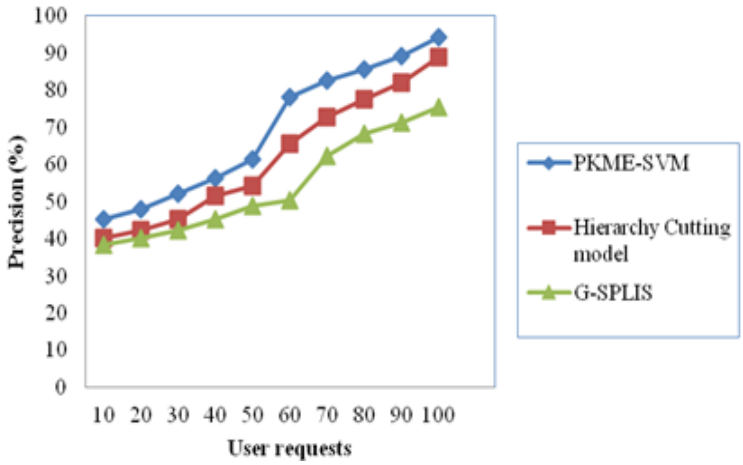

Fig. 6: precision rate of user request retrieval from textual web content.

From figure 6, it is illustrative that the precision ratio is improved using the proposed framework PKME-SVM. For example with 60 user requests, the precision ratio was 78.14 percent using PKME-SVM whereas Hierarchy Cutting model recorded 65.56 percent and 50.32 percent in G-SPLIS. By observing the dense user requests for semantic web mining from textual web content, the precision ratio is improved. This is because with the application of Generalized Association Support Vector Mining algorithm, the precision ratio is increased. With the help of Generalized Association Support Vector Mining algorithm, accurate semantic relation from textual web content are retrieved using latent semantic relationship for mining semantic relation, helps in improving the precision ratio by 12% compared to Hierarchy Cutting model. In addition, by applying dual characteristic function, that obtains optimal term-document matrix that substantially reduces the term redundancy and therefore improves the precision ratio by 28% compared to G-SPLIS.

Conclusion

In this paper, a new Generalized Association Support Vector Mining algorithm is proposed for efficient mining of semantic relation from textual web content with higher accuracy and lesser computation time. The proposed framework is used to balance the sentence and terms generated, increasing the user accuracy retrieval rate to reduce the computation time and precision rate. This work mainly focuses on improving the linear inseparability from textual web content by converting the Probabilistic Term Taxonomy into conceptual graph as well as the no of associative sentences using the PK-Support Vector Machine algorithm. This proposed work has been compared with existing techniques namely Hierarchy

Cutting model and G-SPLIS association semantic and it is proved that the performance of the proposed algorithm is better than the existing algorithms in terms of accuracy, computation time and precision rate.

References

1. Zheng Xu, Shunxiang Zhang, Kim-Kwang Raymond Choo, Lin Mei, Xiao Wei, Xiangfeng Luo, Chuanping Hu, and Yunhuai Liu, "Hierarchy-Cutting Model based Association Semantic for Analyzing Domain Topic on the Web", IEEE Transactions on Industrial Informatics (Volume: 13, Issue: 4, Aug. 2017)
2. Iosif Viktoratos, Athanasios Tsadiras, Nick Bassiliades, "Modeling human daily preferences through a context-aware web-mapping system using semantic technologies", Pervasive and Mobile Computing, Elsevier, Aug 2016
3. Chen Yi, "An English POS Tagging Approach Based on Maximum Entropy", International Conference on Intelligent Transportation, Big Data & Smart City, Jul 2015
4. Diana Maynard, Ian Roberts, Mark A. Greenwood, Dominic Rout, Kalina Bontchev, "Web Semantics: Science, Services and Agents on the World Wide Weba," A framework for real-time semantic social media analysis", Elsevier, Mar 2017
5. Lei Yang and Jianpei Zhang, "Automatic transfer learning for short text Mining", EURASIP Journal on Wireless Communications and Networking, May 2017
6. Rim Rekika, Ilhem Kallel, Jorge Casillas, Adel M. Alimia, "Assessing web sites quality: A systematic literature review by text and association rules mining", International Journal of Information Management, Elsevier, Jan 2018
7. Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge", IEEE Transactions on Knowledge and Data Engineering (Volume: 29, Issue: 3, March 1 2017)
8. Kim Schouten, Onne van der Weijde, Flavius Frasincar, and Rommert Dekker, "Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis With Co-Occurrence Data", IEEE Transactions on Cybernetics (Volume: 48, Issue: 4, April 2018)
9. Jiangchao Yao, Yanfeng Wang, Ya Zhang, Jun Sun, Jun Zhou, "Joint Latent Dirichlet Allocation for Social Tags", IEEE Transactions on Multimedia (Volume: 20, Issue: 1, Jan. 2018)
10. Shu Tian, Xu-Cheng Yin, Ya Su, and Hong-Wei Hao, "A Unified Framework for Tracking Based Text Detection and Recognition from Web Videos", IEEE Transactions on Multimedia (Volume: 20, Issue: 1, Jan. 2018)
11. Ioannis Anagnostopoulos, Manolis Wallace, Sherali Zeadally, "Introduction to the topical issue on semantic social networks and media applications", Social Network Analysis and Mining, Springer, Oct 2017
12. Sunny Sharma1 and Vijay Rana, "Web Personalization through Semantic Annotation System", Advances in Computational Sciences and Technology", Research India Publications, Volume 10, Number 6 (2017)
13. Kaizhu Huang, Rui Zhang, Xiaobo Jin, Amir Hussain, "Special Issue Editorial: Cognitively Inspired Computing for Knowledge Discovery", Cognitive Computation, Springer, Jan 2018-04-05
14. Bissan Audeh, Michel Beigbeder, Antoine Zimmermann, Philippe Jaillon, CeÂdric Bousquet, "Vigi4Med Scraper: A Framework for Web Forum Structured Data Extraction and Semantic Representation", PLOS ONE | DOI:10.1371/journal.pone.0169658 January 25, 2017
15. Halil Kilicoglu, Graciela Rosemblat, Thomas C. Rindflesch, "Assigning factuality values to semantic relations extracted from biomedical research literature", PLOS ONE | https://doi.org/10.1371/journal.pone.0179926 July 5, 2017
16. Kristin Larsson, Simon Baker, Ilona Silins, Yufan Guo, Ulla Stenius, Anna Korhonen, Marika Berglund, "Text mining for improved exposure Assessment", PLOS ONE | DOI:10.1371/journal.pone.0173132 March 3, 2017
17. Gabriel Guerrero-Contreras, José L. Navarro-Galindo, José Samos, and José Luis Garrido, "A Collaborative Semantic Annotation System in Health: Towards a SOA Design for Knowledge Sharing in Ambient Intelligence", Hindawi Mobile Information Systems Volume 2017
18. Yueqin Zhu, Wenwen Zhou, Yang Xu, Ji Liu, and Yongjie Tan, "Intelligent Learning for Knowledge Graph towards Geological Data", Hindawi Scientific Programming Volume 2017
19. Fernando Pech, Alicia Martinez, Hugo Estrada, and Yasmin Hernandez, "Semantic Annotation of Unstructured Documents Using Concepts Similarity", Hindawi Scientific Programming Volume 2017
20. Roberta Akemi Sinoara, João Antunes and Solange Oliveira Rezende, "Text mining and semantics: a systematic mapping study", Journal of the Brazilian Computer Society, Springer, May 2017.