IJCTA, 10(07), 2017, pp. 205-216
© International Science Press

Continuous Point Wise Key Information Extraction For Semantic Data Analysis

Varghese S. Chooralil* and N. Kumar**

Abstract : With the expansion of the web and mobile internet devices people are optimistic to distribute their view on products, news articles, books and any type of subject. It becomes virtually impossible to manually collect significant information and check new trends in an appropriate manner without a mechanized support. In this work, to improve the accuracy and recall rate, a key information extraction method using semantic context from web documents called, Continuous Point-wise Mutual Information Weighting Factor (CPMI-WF) is presented. The method CPMI-WF captures the snippets from the web pages to provide accurate preferences of users for semantic data analysis. First, the key information pre-processing is performed to reduce the noise before key information extraction. Second, the absolute or extracted keywords are used to identify the interesting candidate words. Finally, weighting factor normalization is constructed to minimize the dimensional effects, which maintains the evolution of semantic data analysis. The results obtained through CPMI-WF have been compared with the results obtained from an existing Semantic Data Analysis (SDA) method. The experimental results demonstrate that the proposed method produces significantly higher precision and recall rate than the state-of-the-art works.

Keywords : Point-wise Mutual Information, Weighting Factor, Semantic Data Analysis, Web pages, extracted keywords.

1. INTRODUCTION

The Web is the biggest source of openly available personal opinions, annotations and reviews. The number of new pages available each day on the internet grows exponentially and so does the amount of the obtainable information. At the same time, user profiles were used in web information gathering for the best interpretation of semantic meanings of queries and hence to capture user information needs.

A novel method [1] for providing web page recommendation through semantic enhancement provided an insight into development of web page candidates. With this obtained web page candidates, the method was proved to be better in terms of precision and satisfaction. On the other hand, a personalized ontology model was introduced in [2] for knowledge representation over user profiles with the aid of global and user local information. A technique based on the idea of sparse coding was presented in [3] to improve the predictive performance based on the subspace cluster membership.

With the swift improvement in the video data availability, need to design methods to extract the contents in the video have been occurred. A Fuzzy Ontology and Rule-based model [4] for automatic semantic content extraction was presented that resulted in the satisfactory precision and recall rates to extract object, event. However, it is very difficult to extract from two massively multi core hardware platforms. To address this issue, Term Frequency Inverse Document Frequency (TFIDF) [5] was designed that provided trade off between accuracy and resource utilization.

* Research Scholar,Computer Science & Engineering, Vels University, Chennai-600043. *E-mail: varghesee.kutty@gmail.com*

** Associate Professor,Dept.Of Computer Science & Engineering, Vels University, Chennai. *E-mail: kumar.se@velsuniv.ac.in*

A ranked document list with response to each query is returned in an information retrieval (IR) system. Hence, one of the key issues to be addressed in IR system is the ranking model for the effectiveness of such systems. In [6], Immune Programming based ranking function discovery approach was presented that resulted in the improvement of average precision. However, measures were not taken for composite question analysis. To solve this issue, in [7], a divide and conquer approach was presented that divided the composite question into atomic ones resulting in the improvement of f-core. However, the precision and correctness of answers were not ensured. This was provided by means of combining two types of semantic information in [8].

Semantic expert system attempts to process and understand natural language that refers to complex and diverse context of human language. In the recent era, the Natural Language Processing (NLP) field presents a requirement for efficient algorithms and techniques to measure the text and sentence similarity. A new sentence similarity measure via semantic vector space was introduced in [9] that resulted in the semantic similarity of sentence extraction provided with arbitrary structures. Another novel method based on proximity measure was designed in [10] that used semantic and statistical information. A portable form-based search interface for semantic datasets was presented in [11] using multiclass search.

With the increasing gain of symbol-like processing, many research works have been contributed in this area. In [12], a Neural Engineering Framework (NEF) was designed for optimizing the semantic pointer representations. Semantic factors prediction of lexical replacement estimates was provided in [13] through linear regression model.

Although certain methods have been proposed for each of the above-mentioned approaches (semantic content extraction, extraction from multi core hardware platforms, composite question analysis and optimizing semantic representations), semantic data analysis stays a largely unexplored research avenue. In this paper, we concentrate on improving the precision and reducing the computational complexity (*i.e*. key frame extraction time) based on structured keyword extraction.

This research paper presents a method for semantic data analysis and Continuous Point-wise Mutual Information Weighting Factor (CPMI-WF) using two-way analysis by investigating extracted keywords. The main focus is on improving the precision, recall rate and reducing the key frame extraction time for semantic data analysis.

2. RELATED WORKS

Real-time data processing from multiple heterogeneous data streams and static databases is considered to be one of the most typical tasks in several industrial scenarios. To perform complex types of diagnostic, hundreds of queries have to be collected to obtain relevant information. Besides, several amounts of queries retrieve data of the same kind, but structurally different sources are accessed.

In [14], Semantic Technologies were implemented to simplify the task of complex diagnostics by way of designing an abstraction layer, called ontology in order to integrate the heterogeneous data. However, information regarding specific aspect remained unaddressed. To solve this issue, in [15], a finer-grained problem of aspect-oriented mining was presented that not only addressed the aspect-oriented part, but also improved the predictions of aspect-oriented opinions. Besides, improvement observed in accuracy, the error rate at which the accuracy was provided remained unaddressed. A Semi-Automated Text Classification [16] based on the notion of inspection gain was provided that resulted in the reduction in classification errors.

Researchers in the field of life sciences perform scientific literature search as part of their daily routines and activities. In [17], a semantic ranking and result visualization system was presented in order to find relevant references in a more interactive and easy manner. To measure semantic similarity for geo-knowledge graphs, Network-Lexical Similarity (NLS) measure was obtained in [18] with higher correlation rate. A semantic similarity ensemble was presented in [19].

Following the semantic analysis methods presented above, we propose a new continuous point-wise mutual information weighting factor-based semantic data analysis to reduce the key frame extraction time and improve the precision recall rate for semantic content extraction. The experimental result shows how our method achieves high precision rate and efficient performance.

3. CONTINUOUS POINT-WISE MUTUAL INFORMATION WEIGHTING FACTOR

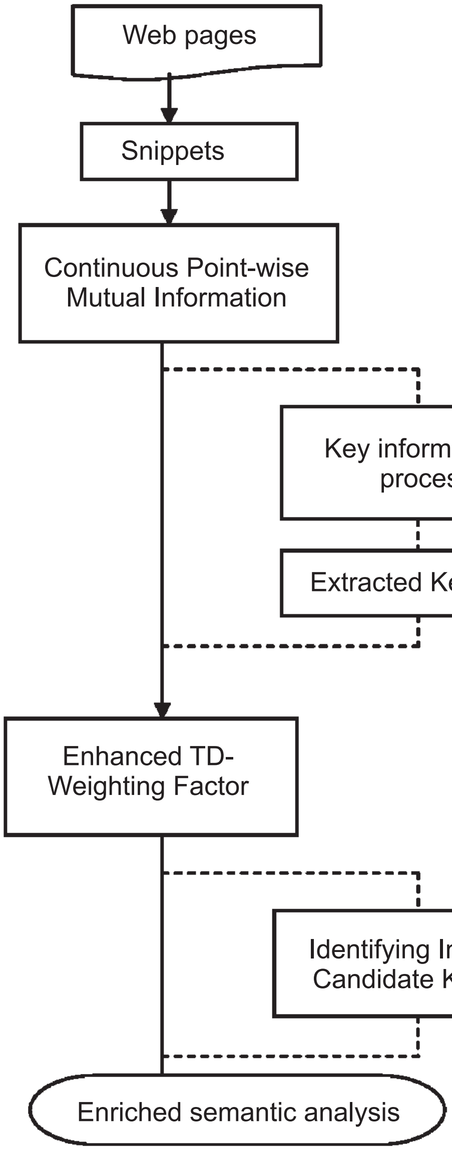
Web pages
Snippets
Continuous Point-wise Mutual Information
Key information pre-processing
Extracted Keywords
Enhanced TD-Weighting Factor
Identifying Interesting Candidate Keywords
Enriched semantic analysis

Figure 1: Illustration of CPMI-WF based Semantic Data Analysis

In this section, Continuous Point wise Mutual Information Weighting Factor for structured key information extraction is explained as below. First of all the web page content is fetched in the form of web snippets for extracting key information along with the search results. At a second step Term Frequency and Weighting Factor is applied to each extracted keyword with the objective of identifying the interested candidates, the desired metadata, for semantic data analysis. Fig. 1 shows the illustration of Continuous Point-wise Mutual Information Weighting Factor-based (CPMI-WF) Semantic Data Analysis.

The framework of Continuous Point-wise Mutual Information Weighting Factor (CPMI-WF) for semantic data analysis shown in the above figure consists of two steps namely, Key Information Pre-processing using Continuous Point-wise Mutual Information and Identifying Interesting Candidate Keywords by applying Enhanced TD-Weighting Factor. The CPMI applied in the CPMI-WF method reduces the key frame extraction time during pre-processing and therefore significantly results in the improvement of recall rate. By applying ETD-WF in the CPMI-WF method, there results in the improvement of precision and semantic data analysis efficiency. The elaborate description of CPMI-WF based semantic data analysis is provided in the following sections

A. Key Information Pre-processing using Continuous Point-wise Mutual Information

The first step in the CPMI-WF method is the key information pre-processing using Continuous Point-wise Mutual Information (CPMI). As several stop words occur on snippets, preprocessing of the snippets are highly required to minimize the noise before extracting the key information. To reduce the noise, the CPMI-WF method applies CPMI and is illustrated in the figure.

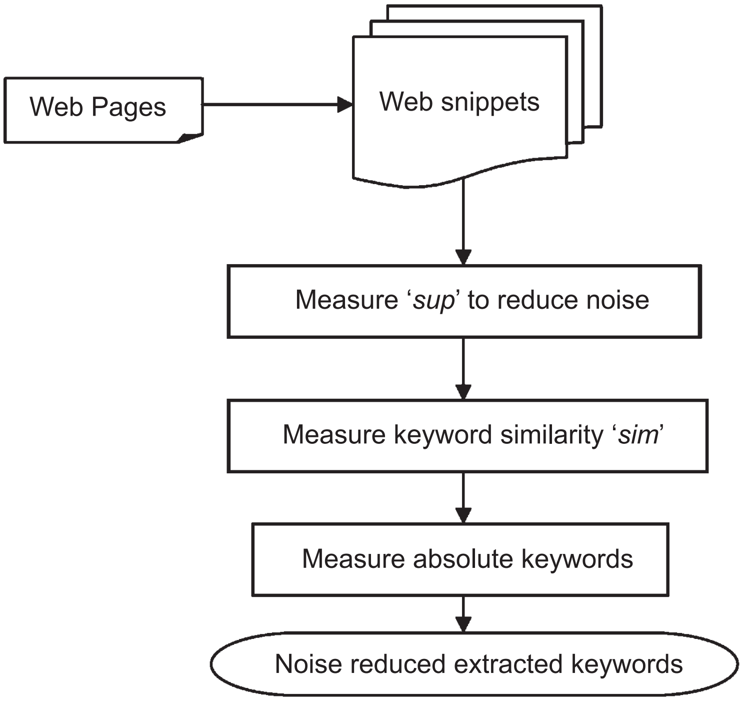
Web Pages
Web snippets
Measure '*sup*' to reduce noise
Measure keyword similarity '*sim*'
Measure absolute keywords
Noise reduced extracted keywords

Figure 2: Illustration of Key Information Pre-processing using Continuous Point-wise Mutual Information

As shown in the figure, key information pre-processing is performed using CPMI. If a keyword occurs more than once in a snippet, then it represent an important keyword related to the web page '*wp*' as it coincides in close presence with the web page. Then, the support '*sup*' for extracting the keyword 'k_i' with respect to the returned snippets arise from the web page '*wp*' is formulated as given below.

$$\text{sup}(k_i) = \frac{sr\,(k_i)}{n} \tag{1}$$

From (1), key extraction support value '*sup*' is obtained with the aid of the recurrence '*r*' of snippets '*s*' with respect to the corresponding keyword 'k_i' to the '*n*' number of snippets returned by the web page '*wp*'. Besides the support of keyword 'k_i', relations between keywords are also obtained. Continuous Point-wise Mutual Information is applied to measure the relation between the keywords 'k_i' and 'k_j' respectively. It is mathematically formulated as given below.

$$\text{sim}\,(k_i, k_j) = \frac{\log\,(n{*}sr\,(k_i \cap k_j))/\,sr\,(k_i) * sr(k_j)}{\log n} \tag{2}$$

From (2), '$sr\,(k_i \cap k_j)$', represents the joint snippet recurrence of the keywords 'k_i', and 'k_j', with the normalization factor being 'log *n*' respectively. With the joint snippet recurrence based on the keyword similarity, the CPMI-WF method uses keyword vector representation to obtain the absolute keywords. The absolute keywords refer to the keywords that occur in the snippets (*i.e*. extracted keywords after preprocessing) of the web pages.

Therefore, the keyword that occurs in the snippets from web pages is considered as the absolute keywords of the user domain. Let us consider a scenario when the web user searches the query ‘*Bravia*’, then the snippet contains the keyword ‘*Sony*’ and thus the keyword ‘*Sony*’ is the absolute keyword (*i.e*. extracted keyword) of the user. Hence, the user domain ‘*ud*’ in the CPMI-WF method is represented in the form of keyword vector and is mathematically formulated as given below.

$$ek \rightarrow ud \;=\; \{w\,(1, ud), w\,(2, ud), \ldots\, w(n, ud)\} \tag{3}$$

From (3), ‘*w*’ refers to the weight of the ‘*ith*’ keyword in the user domain ‘*ud*’ that is stored as the resultant noise reduced extracted keyword ‘*ek*’.

$$w\,(i, ek) \;=\; \begin{cases} 1, \text{if } k_i \in s_j \\ 0, \text{Otherwise} \end{cases} \tag{4}$$

From (4), based on the weight measure, upon clicking the snippet ‘s_j’, the weight of keyword ‘k_i’ is set to ‘1’, otherwise is set to ‘0’. Figure 3 shows the Continuous Point-wise Mutual Information algorithm for extracting noise reduced keywords.

As illustrated in the algorithm (fig. 3), the main objective in the CPMI-WF for web analysis and text classification lies in pre-processing the web pages and allocating high level data types to the content. This is attained through evaluating the joint snippet recurrence and obtaining the absolute keywords through it, on the basis of the user domain.

Input : Web Page ‘*wp*’, Snippets ‘S = $s_1, s_2, \ldots, s_n$’, user domain ‘*ud*’

Output : Noise reduced keywords

```
Begin
    For each Web Page ‘wp’
        For each Snippets ‘S’
            For each user domain ‘ud’
                Measure key extraction support value using (1)
                Measure joint snippet recurrence using (2)
                Measure absolute keywords using (3)
            End for
        End for
    End for
End
```

Figure 3: Continuous Point-wise Mutual Information algorithm

B. Identifying Interesting Candidate Keywords using Term Frequency and Weighting

Semantic data analysis has distinguishing feature and in the proposed method Trip Advisor dataset is used for analysis. Automatically Extracted Keyword (AEK) is designed in the CPMI-WF method to extract interested candidates which include the Term Frequency (TF) and Weighting, and the Enhanced TF-Weighting Factor algorithms. The Term Frequency [5] and Weighting reflects the significance and semantic data analysis of the extracted keyword in a web page. The main part of TF is that if a keyword has high frequency in particular snippets in a web page, then this type of keyword bears high capacity to distinguish category, and thus possessing high degree of significance. The TF for the extracted keyword then refers to the frequency, the extracted keyword occurs in a snippet. It is mathematically formulated as given below.

$$\text{TF}\,(ek, s_i) \;=\; \sum_{i=1}^{n} \frac{\text{Count}\,(ek, s_i)}{\text{Count}\,(ek_i, s_i)} \tag{5}$$

From (5), the frequency of extracted keyword ‘*ek*’ in a snippet ‘s_i’ is obtained, whereas ‘ek_i’ refers to all extracted keywords. With the obtained Term Frequency for the extracted keyword, the CPMI-WF method then measures the semantic similarity between them. The CPMI-WF method extracts the features of two keywords in any web page based on the semantic similarity using the cosine method. The semantic similarity using the cosine method is as given below.

$$\cos(ek_i, ek_j) = \frac{\sum_{i=1}^{bow} A_i B_i}{\sqrt{\sum_{i=1}^{bow} A_i}\sqrt{\sum_{i=1}^{bow} B_i}} \tag{6}$$

From (6), 'ek_i' and 'ek_j' represents the two extracted keywords, whereas 'A_i' and 'B_i' represents the words from bag of words '*bow*' respectively. Followed by this, the weighting factor with the extracted keywords is obtained for the Trip advisor dataset. This is performed with the help of an illustration as shown in the figure 4.

Let us consider a dataset called, Trip Advisor that comprises of reviews randomly selected from several accommodations. In order to obtain the weighting factor, two snippets, *i.e.*, room file snippet and value file snippets are considered and is mathematically formulated as given below.

$$T_1 = \frac{\max_{i=1,2,\ldots N_1}(\text{Count}(ek, c_i))}{\text{Count}(ek, c_1))} \tag{7}$$

$$T_2 = \frac{\max_{i=1,2,\ldots N_2}(\text{Count}(ek, c_i))}{\text{Count}(ek, c_2))} \tag{8}$$

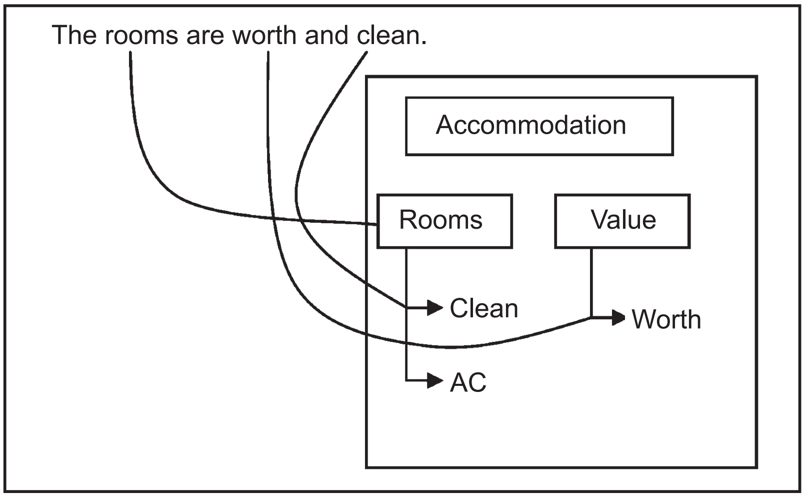

Figure 4: Illustration for measuring weighting factor using Trip Advisor dataset

From (7), the weighting factor for the extracted keyword 'T_1' is measured using the room file snippets 'C_1' and total number of room file snippets 'N_1' from Trip advisor dataset. In a similar manner, from (8), the weighting factor for the extracted keyword 'T_2' is measured using the value file snippets 'C_2' and total number of value file snippets 'N_2' from Trip advisor dataset. Due to different dimension and dimensional units, the CPMI-WF method applies the MAX-MIN weighting factor normalization. The MAX-MIN weighting factor normalization is mathematically formulated as given below.

$$\beta = \frac{T_i - \text{MIN}(T_i)}{\text{MAX}(T_i) - \text{MIN}(T_i)} \tag{9}$$

With MAX-MIN weighting factor normalization 'β' obtained through (9), the Enhanced TD-Weighting Factor is measured as given below.

$$\text{ETDWF} = \beta * \text{TF}(ek, S_i) \tag{10}$$

From (10), 'S_i' represents the snippets collection numbered '*i*', 'TF' represents the extracted keywords '*ek*' frequency in 'S_i', whereas 'β' corresponds to the normalization factor used for weighting. Fig. 5 illustrates the Semantic Data Analysis using Enhanced TD-Weighting Factor.

```
Input : Extracted Keywords '$ek = ek_1, ek_2, ...ek_n$', weighting factor normalization '$\beta$', room file snippets '$C_1$', total number of room file snippets '$N_1$', value file snippets '$C_2$', total number of value file snippets '$N_2$'
Output : Improved semantic data analysis efficiency
Begin
    For each Extracted Keywords '$ek$'
        Measure Term Frequency for the Extracted Keyword using (5)
        Measure cosine semantic similarity using (6)
        Evaluate MAX-MIN weighting factor normalization using (9)
        Measure Enhanced TD-Weighting Factor using (10)
    End for
End
```

Figure 5: Enhanced TD-Weighting Factor-based Semantic Data Analysis

As shown in the fig. 5, for each extracted keywords, semantic data analysis is performed by obtaining the Enhanced TD-Weighted Factor. To evolve this, the algorithm first obtains the term frequency with the extracted keyword. Followed by this, the semantic similarity for the extracted keyword is obtained by measuring the cosine similarity function. Next, normalization is performed using the MAX-MIN weighting factor. Finally the product of term frequency and normalized weighting factor is used to extract the semantic data for analysis.

4. EXPERIMENTAL SETUP

In this section, the results of a series of experiments carried out to evaluate the effectiveness of the proposed method and to compare with other state-of-the-art methods are presented. Continuous Point-wise Mutual Information Weighting Factor (CPMI-WF) for Semantic Data Analysis uses JAVA language with Weka tool for the experimental work. The CPMI-WF method uses the TripAdvisor Dataset from TripAdvisor.com for the experimental work.

The TripAdvisor Dataset contains 200 reviews from TripAdvisor.com randomly selected from several accommodations. It covers all 5 satisfaction levels (40 reviews in each level) consisting of 1,382 criticisms, 211 non-criticisms, and 97 criticisms with errors. The information is collected from the Tripadvisor and Edmunds. The Tripadisor shows the 259000 reviews. The experiment is conducted on the factors such as number of reviews, key frame extraction time, recall rate, precision and semantic data analysis efficiency. In order to evaluate the performance of the CPMI-WF method, certain metrics are introduced to measure the semantic data analysis and compared with the existing methods namely, Domain Ontology of Web Pages (DOWP) [1] and Proposed Ontology (PO) [2] model. The performance metrics are in the following.

A. Key frame extraction time demonstration of CPMI-WF

The main goal of our experiments is to determine the key frame extraction time for semantic data analysis. We randomly selected a review set of about 105 from Trip Advisor.com. With this experimental setting the rate of key frame extraction time is defined as given below.

The key frame extraction time is the time taken to extract the key frames (*i.e*. extracted keys) with respect to the total number of reviews in web pages. The Key frame extraction time is measured as given below.

$$KFE_t = r_i * \text{Time}\ (ek) \tag{11}$$

From (11), 'KFE_t' refers to the key frame extraction time using the number of reviews 'r_i' for the extracted keywords '*ek*' respectively. It is measured in terms of milliseconds (*ms*). Lower key frame extraction time signifies the efficiency of the method presented. Table 1 presents the results of key frame extraction time of an exploratory experimentation on Trip Advisor dataset by presenting the key frame extraction time using CPMI-WF, DOWP and PO. The experiment was conducted to gain insights on the semantic data analysis of the datasets, to measure the performance of key frame extraction time with respect to 105 different reviews.

Table 1

Results of Key Frame Extraction Time

No. of reviews	*Key frame extraction time (ms)*		
	CPMI-WF	*DOWP*	*PO*
15	6.82	8.3	9.5
30	10.43	12.54	14.64
45	15.32	17.43	19.53
60	22.13	24.24	26.34
75	28.98	30.16	32.26
90	34.14	36.25	38.35
105	40.28	42.39	44.49

For all scenarios as shown in table 1, key frame extraction time is increasing with total number of reviews obtained from different users. Seven unique experiments were conducted for each review size.

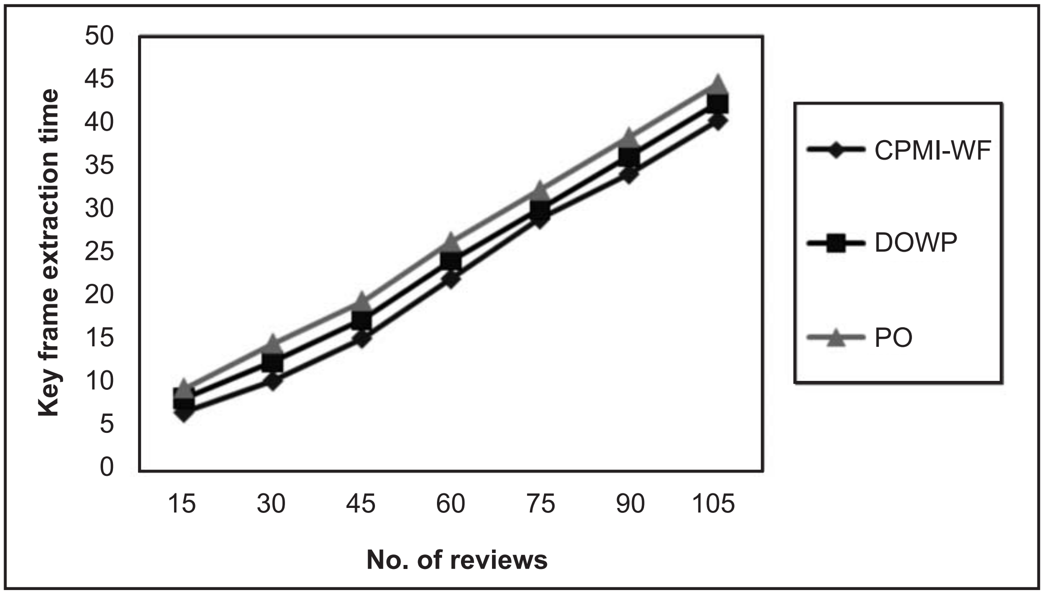

Figure 6: Measure of key frame extraction time

The results presented in fig. 6 shows the key frame extraction time when users are presented with several reviews. From the fig. 6 it is clear that the values of key frame extraction time increase with the increase in the number of reviews using all the three methods CPMI-WF, DOWP and PO respectively. The important observation from the figure is that the key frame extraction time is directly proportional to the number of review. Therefore though major deviations are not being observed, but comparatively the CPMI-WF proved to be better. This is because of the application of Continuous Point-wise Mutual Information that performs the process of pre-processing based on the support of keywords being extracted and their similarity observed through the cosine function. This in turn based on the weight measure obtains the absolute keywords and therefore reduced the key frame extraction time of CPMI-WF by 12% compared to DOWP and 23% compared to PO. In particular, it shows an average improvement of 17% over the original approach in the extraction of key frame.

B. Precision rate demonstration of CPMI-WF

In the experiment, to clearly compare the features of both CPMI-WF and existing Domain Ontology of Web Pages (DOWP) [1] and Proposed Ontology (PO) [2] model, we simplify the precision rate involved in semantic data analysis as following defined. Precision rate refers to the number of relevant extracted keywords with respect to the number of returned keywords or it refers to the number of relevant returned extracted keywords of the web page '*wp*' with regard to the 'N' returned extracted keywords.

$$\mathrm{P} \;=\; \sum_{i=1}^{n}\left(\frac{\mathrm{Rel}\,(ek_i)}{\mathrm{N}}\right) * 100 \tag{12}$$

From (12), the precision rate 'P' is obtained, using the relevant extracted keywords, 'Rel(ek_i)' and the total number of extracted keywords, 'N' from web pages. The results of precision rate for different semantic data analysis model are provided in Table 2.

Table 2
The Results of Precision Rate

No. of reviews	*Precision rate (%)*		
	CPMI-WF	*DOWP*	*PO*
15	73.45	65.89	61.28
30	70.25	62.25	58.22
45	68.49	60.47	56.44
60	66.32	58.30	54.27
75	64.19	56.17	52.14
90	78.93	70.91	66.87
105	85.89	78.87	74.83

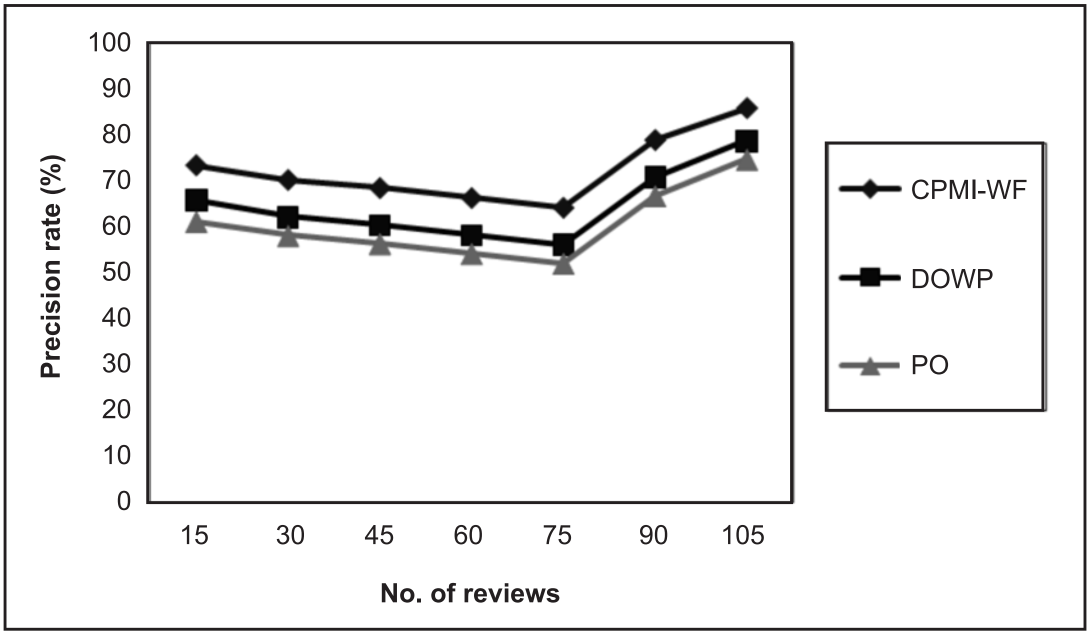

Figure 7: Measure of precision rate

Results from fig. 7 indicate that precision rate for CPMI-WF is greater than DOWP and PO. Furthermore, the improvement in the precision rate performance achieved by CPMI-WF method over the existing DOWP and PO was compared to be higher. This is because of the application of Term Frequency where optimal solutions are obtained due to the updated keywords, i.e. extracted keywords (not the actual keywords before preprocessing). This in turn confirms the improved precision rate for semantic data analysis by applying CPMI-WF than DOWP and PO. Another interesting observation from figure is that CPMI-WF was capable of differentiating the actual keywords and extracted keywords from the overall reviews by measuring the semantic similarity using the cosine method. Baseline results were lower than 70% when using 15 and 30 labeled examples, indicating that the initial selection of the reviews are very difficult to analyze. On the other hand, the upper-bound with increasing number of reviews saw a good result by improving the precision rate by 11% compared to DOWP and 17% compared to PO respectively.

C. Recall rate demonstration of CPMI-WF

Finally, we address the third goal of the experiments with respect to recall rate for semantic data analysis showing the comparison between CPMI-WF, DOWP and PO and defined as follows. Recall rate refers to the number of relevant extracted keywords with respect to the number of relevant keywords or it refers to the number of relevant returned extracted keywords of the web page '*wp*' with regard to the 'Rel(*k*)' returned relevant keywords.

$$P = \sum_{i=1}^{n} \left(\frac{\text{Rel}\,(ek_i)}{\text{Rel}\,(k)} \right) * 100 \qquad (13)$$

From (13), the recall rate 'R' is obtained, using the relevant extracted keywords, 'Rel(ek_i)' and the total number of relevant keywords, '*k*' from web pages.

Table 3
Results of Recall Rate

No. of reviews	*Recall rate (%)*		
	CPMI-WF	*DOWP*	*PO*
15	92.35	81.48	76.29
30	88.15	77.26	71.23
45	85.28	74.39	69.36
60	77.32	66.43	61.40
75	81.28	70.39	65.36
90	85.89	74.22	69.19
105	93.42	82.53	77.50

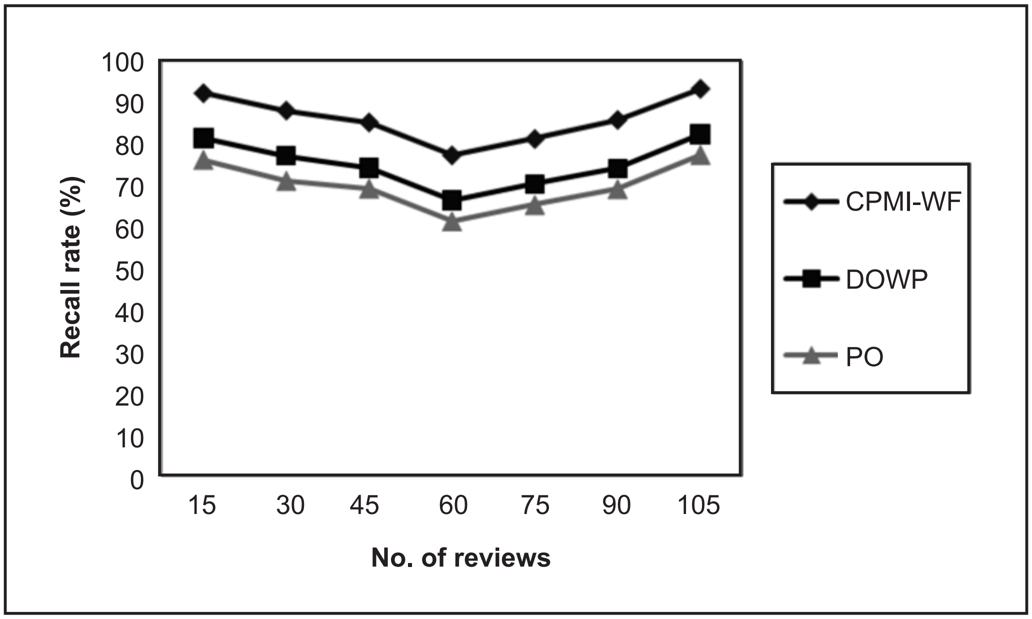

Figure 8: Measure of recall rate

Table 3 and Fig. 8 show the recall rate semantic data analysis accuracy versus number of reviews provided. The experiments were conducted with different number of reviews in the range of 15 to 105. Results from figure presents the reported recall rate involved in the semantic data analysis. For this kind of reviews the best result of the proposed method was 'r = 93.42%' using '105 instances (reviews)' for training. In contrast, the existing method achieved '89.53% and 77.50' using DOWP and PO respectively. This is because of applying the Continuous Point-wise Mutual Information algorithm joint snippet recurrence based on cosine similarity was obtained. With this measure, the semantic similarity between the extracted keywords was obtained. This in turn resulted in the improvement of recall rate of CPMI-WF by 13% compared to DOWP. In addition, using Enhanced TD-Weighting Factor-based Semantic Data Analysis, relevant extracted keywords were obtained at reduced time interval and therefore improving the recall rate of CPMI-WF by 19% compared to PO respectively.

5. CONCLUSION

We have developed Continuous Point-wise Mutual Information Weighting Factor-based semantic data analysis that extracts the key information with high performance from web pages. The method decomposes the web pages into snippets. The method initially focuses on extracting the keywords based on Continuous Point-wise Mutual Information and obtaining absolute keywords using support and cosine similarity function. The method has improved the recall rate, since it has used the Continuous Point-wise Mutual Information algorithm to retrieve all relevant extracted keywords with respect to the relevant keywords. The method Enhanced TD-Weighting Factor algorithm by using MAX-MIN weighting factor normalization improves the semantic data analysis efficiency. Experimental result indicates that the CPMI-WF based semantic data analysis outperforms all the existing semantic data analysis work with 16% improved recall rate and 14% improvement in precision rate.

6. REFERENCE

1. Thi Thanh Sang Nguyen, Hai Yan Lu, Jie Lu, "Web-page Recommendation based on Web Usage and Domain Knowledge", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 10, October 2014, Pages 2574 – 2587.
2. Xiaohui Tao, Yuefeng Li, and Ning Zhong, "A Personalized Ontology Model for Web Information Gathering", IEEE Transactions On Knowledge And Data Engineering, Volume 23, Issue 4, April 2011, Pages 496-511.
3. Brian Quanz, Jun (Luke) Huan, and Meenakshi Mishra, "Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue", IEEE Transactions On Knowledge And Data Engineering, Volume 24, Issue 10, October 2012, Pages 1789-1802.
4. Yakup Yildirim, Adnan Yazici and Turgay Yilmaz, "Automatic Semantic Content Extraction in Videos Using a Fuzzy Ontology and Rule-Based Model", IEEE Transactions On Knowledge And Data Engineering, Volume 25, Issue 1, January 2013, Pages 47-61.
5. Craig Ulmer, Maya Gokhale, Brian Gallagher , Philip Topa, Tina Eliassi-Rad , "Massively parallel acceleration of a document-similarity classifier to detect web attacks", Elsevier, Journal of Parallel and Distributed Computing , July 2011, Pages 225–235.
6. Shuaiqiang Wang, Jun Ma, Qiang He, "An immune programming-based ranking function discovery approach for effective information retrieval", Elsevier, Expert Systems with Applications, Volume 37, Issue 8, August 2010, Pages 5863–5871.
7. Hyo-Jung Oh a, Ki-Youn Sung c, Myung-Gil Jang a, Sung Hyon Myaeng, "Compositional question answering: A divide and conquer approach", Elsevier, Information Processing & Management, Volume 47, Issue 6, November 2011, Pages 808–824.
8. Paloma Moreda , Hector Llorens, Estela Saquete, Manuel Palomar, "Combining semantic information in question answering systems", Elsevier, Information Processing & Management, Volume 47, Issue 6, November 2011, Pages 870–885.
9. Ming Che Lee, "A novel sentence similarity measure for semantic-based expert systems", Elsevier, Expert Systems with Applications, Volume 38, Issue 5, May 2011, Pages 6392–6399.

10. Liu Wenyin, Xiaojun Quan, Min Feng, Bite Qiu, “A short text modeling method combining semantic and statistical information”, Elsevier, Information Sciences, Volume 180, Issue 20, 15 October 2010, Pages 4031–4041.

11. Guillermo Vega-Gorgojo, Martin Giese, Simen Heggestøyl, Ahmet Soylu, Arild Waaler, “PepeSearch: Semantic Data for the Masses”, Plos one, Research Article, March 2016, Pages 1-1

12. Jan Gosmann, Chris Eliasmith, “Optimizing Semantic Pointer Representations for Symbol-Like Processing in Spiking Neural Networks”, Plos one, Research Article, February 2016, Pages 1-18.

13. Susanne Vejdemo, Thomas Hörberg,“Semantic Factors Predict the Rate of Lexical Replacement of Content Words”, Plos one, Research Article, January 2016, Pages 1-15.

14. E. Kharlamov, S.Brandt, M.Giese, E.Jiménez-Ruiz, Y.Kotidis S. Lamparter T.Mailis C. Neuenstadt, Ö.Özçep, C.Pinkel, A.Soylu, C.Svingos, D.heleznyakov, I.Horrocks, Y.Ioannidis, R.Möller, A.Waaler, “Demo: Enabling Semantic Access to Static and Streaming Distributed Data with Optique”, ACM International Conference on Distributed and Event-based Systems, June 2016, Pages 350-353.

15. Diego Marcheggiani, Oscar Tackstrom, Andrea Esuli, and Fabrizio Sebastiani, “Hierarchical Multi-label Conditional Random Fields for Aspect-Oriented Opinion Mining”, Springer, International Publishing Switzerland, April 2014, Pages 273-285.

16. Giacomo Berardi, Andrea Esuli, and Fabrizio Sebastiani, “A Utility-Theoretic Ranking Method for Semi-Automated Text Classification”, International ACM SIGIR conference on Research and development in information retrieval, Volume 10, Issue 1, July 2015, Pages 961-970.

17. Julia Stoyanovich, Mayur Lodha, William Mee, Kenneth A. Ross, “SkylineSearch: Semantic Ranking and Result Visualization for PubMed”, ACM SIGMOD International Conference on Management of data, June 2011, Pages 1247-1250.

18. Andrea Ballatore , Michela Bertolotto and David C. Wilson, “A Structural-Lexical Measure of Semantic Similarity for Geo-Knowledge Graphs”, ISPRS International Journal of Geo-Information, April 2015, Pages 471-492.

19. Andrea Ballatore, Michela Bertolotto, and David C. Wilson, “The semantic similarity ensemble”, Journal Of Spatial Information Science, August 2013, Pages 27–44.