Computers, Materials & Continua
DOI:10.32604/cmc.2021.012796
Article

Tech Science Press

A Combinatorial Optimized Knapsack Linear Space for Information Retrieval

Varghese S. Chooralil[1], Vinodh P. Vijayan[2], Biju Paul[1], M. M. Anishin Raj[3], B. Karthikeyan[4,*] and G. Manikandan[4]

[1]Rajagiri School of Engineering & Technology, Kochi, 682039, India
[2]Mangalam College of Engineering, Kottayam, 686631, India
[3]Department of CSE, Viswajyothi College of Engineering & Technology, Vazhakulam, 686670, India
[4]School of Computing, SASTRA Deemed To Be University, Thanjavur, 613401, India
[*]Corresponding Author: B. Karthikeyan. Email: bkarthikeyan@it.sastra.edu
Received: 12 July 2020; Accepted: 17 October 2020

Abstract: Key information extraction can reduce the dimensional effects while evaluating the correct preferences of users during semantic data analysis. Currently, the classifiers are used to maximize the performance of web-page recommendation in terms of precision and satisfaction. The recent method disambiguates contextual sentiment using conceptual prediction with robustness, however the conceptual prediction method is not able to yield the optimal solution. Context-dependent terms are primarily evaluated by constructing linear space of context features, presuming that if the terms come together in certain consumer-related reviews, they are semantically reliant. Moreover, the more frequently they coexist, the greater the semantic dependency is. However, the influence of the terms that coexist with each other can be part of the frequency of the terms of their semantic dependence, as they are non-integrative and their individual meaning cannot be derived. In this work, we consider the strength of a term and the influence of a term as a combinatorial optimization, called Combinatorial Optimized Linear Space Knapsack for Information Retrieval (COLSK-IR). The COLSK-IR is considered as a knapsack problem with the total weight being the "term influence" or "influence of term" and the total value being the "term frequency" or "frequency of term" for semantic data analysis. The method, by which the term influence and the term frequency are considered to identify the optimal solutions, is called combinatorial optimizations. Thus, we choose the knapsack for performing an integer programming problem and perform multiple experiments using the linear space through combinatorial optimization to identify the possible optimum solutions. It is evident from our experimental results that the COLSK-IR provides better results than previous methods to detect strongly dependent snippets with minimum ambiguity that are related to inter-sentential context during semantic data analysis.

Keywords: Key information extraction; web-page; context-dependent; non-integrative; combinatorial optimization; knapsack

1 Introduction

Due to the wide popularity of user reviews in online media, a vast amount of content has been generated over the past several years. An approach to disambiguate the context-based sentiment polarity of words, as an information recovery problem was presented in [1,2]. The recommendation of web pages plays a major role in the web space. Better web page recommendations can be provided through semantic enhancement, as presented in [1–3]. The evaluation of information technology retrieval plays a crucial role in adjudicating documents. A multi-armed bandit model is presented in [4] for a pooling-based evaluation, which will minimize the assessment effort. However, the increase in the reviews has resulted in a substantial reduction of the hit rate. To improve the hit rate, a machine learning approach was presented in [5,6]. A box clustering segmentation model is presented in [7,8] using a clustering algorithm as an accurate reference algorithm.

Semantic data analysis is a field of study in which specific data in a particular domain are analyzed by inputting a query from the search engine. Existing applications have shown that there is vast market potential for semantic data analysis [9] and that the knowledge extracted from users remains the key in many sectors of the society. Multilingual semantic analysis has provided insight into emotional classification, resulting in the improvement in classification performance.

Multiple stages of semantic composition for context-sensitive scalar objectives using the time window model is presented in [10], which shows semantic improvement processing. Another restrictive vs. non-restrictive nominal modification model based on prenominal adjectives was investigated in [11]. A potential study of negative polarity sensitivity is designed in [12] through semantic assertion.

The main goal of this work is to build up a combinatorial optimization method considering inter-sentential context at the bottom level of granularity using linear space with a knapsack called Combinatorial Optimized Linear Space Knapsack for Information Retrieval (COLSK-IR). Instead of relying on snippet and manually labeled datasets to capture diverse kinds of non-integrative terms, the planned method suggests an individual snippet influence term and a query influence by using a combinatorial factor determination.

2 Related Works

Key information extraction is a fundamental technique in the evaluation of information retrieval evaluation and has attracted attention for decades. Based on news corpora, multi-word expression extraction using context analysis and model-based analysis is provided in [13]. While keyword query warrants conventional users to explore an enormous amount of data, the ambiguity of keyword query makes it difficult to efficiently answer keyword queries, specifically for short and vague keyword queries.

In [14] XML keyword search diversification model is presented to improve the precision of query diversification. However, the XML keyword search diversification model does not generally work well for long and complex queries. To address this issue, a key concept identification approach is explored in [15] to improve the query retrieval rate.

Query facets provide us with essential knowledge related to a query and hence are used to enhance the search experience in several ways. An automatic mining model through extraction and grouping of frequent lists is presented in [16], resulting in the mining of better query facets. A new optimized Monte Carlo algorithm is designed in [17] to significantly reduce the number of iterations and computational complexity.

Another graph-based approach to build automatically a taxonomy, resulting in the maximization of the overall associative strength is presented in [18]. A semantic-based, analysis architecture to explore more complex semantic data models based on a case study in commodity pricing is investigated in [19]. The advancement of semantics is an important research area which is significantly challenged by the lack of

ubiquitous metrics to address precision and abnormality pertinent to each domain [20]. However, the search efficiency was compromised with domain-independent text and structural similarity measures.

To enhance the efficiency of latent semantic models in web search, meta-features are created in [21], which uses feature vectors. With a feature vector, a model's forecast for a given query document pair is then passed to the overall ranker in addition to the models' scores. This in turn results in improved performance of latent semantic models. A language-independent framework to retrieve high precision queries using the traditional bootstrapping approach is presented in [22]. A Context Aware Time Model (CATM) in [23] provides an insight to the user actions at varying time intervals.

Our study covers both the detection of strongly dependent snippets and the reduction in ambiguity related to inter-sentential context to test whether the sarcastic use of the word has an influential factor in the COLSK-IR method. The work also covers the knapsack-based combinatorial optimization for semantic data analysis as a possible way to obtain an evidence for an effective semantic linear space representation.

3 Combinatorial Optimized Linear Space Knapsack for Information Retrieval

The contextual polarity of a word [6] is taken into consideration by many factors. For example, it could be difficult to detect a sarcastic use of the word "great" in the sentence "That's great!" without considering [24,25] inter-sentential context. With respect to this lack of difference between snippet, term, influence and query influence, this paper presents a combinatorial optimization method using linear space with knapsack, called Combinatorial Optimized Linear Space Knapsack for Information Retrieval (COLSK-IR).

The basic idea behind the COLSK-IR method is presented with a set of items, where weight and value are available for all. The combinatorial optimization model measures the number of item to be included in a set so that the calculated weight is always below or same as the given limit and the total value is as large as possible. The block diagram of COLSK-IR is shown in Fig. 1. The initial process starts with the fetching of web snippets from the web page content to meet the criteria for obtaining an optimal solution. This method detects non-integrative queries using total weight and total value as a combinatorial optimization problem.

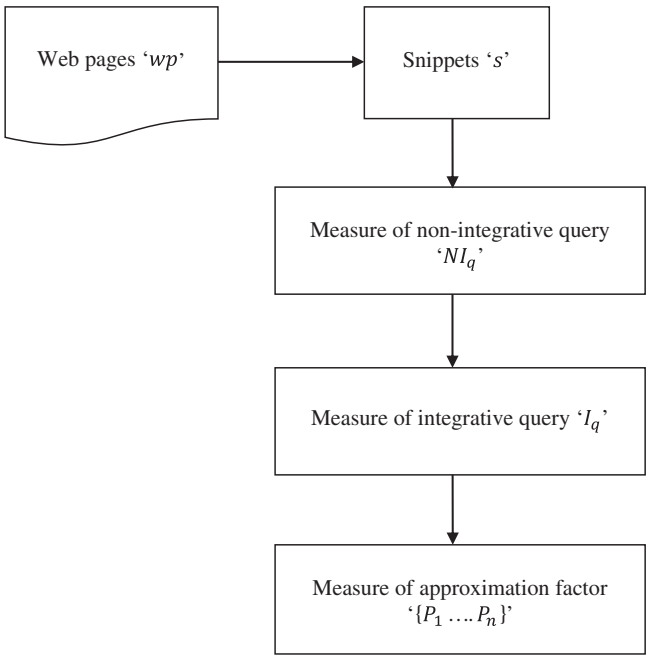

Figure 1: Block diagram of COLSK-IR

3.1 Non-Integrative Perturbed Approximation

The basic COLSK-IR method consists of substituting a keyword 'k_i' from a web page 'wp' by a synonym 'S_j' and measuring the semantic separation '|' of the replacement keyword '$k_i^{'}$' from the novel snippet 's.' If their meanings differ, it is more likely that the original snippet is non-integrative. However, if their meanings do not differ, then the original snippet is less likely to be non-integrative. We express this perturbation as follows. Let 'wp' represent the web page of query 'q' containing prearranged set of snippets or terms 'n', where 's' is the snippet conveyed from 'n' number of snippets.

$$wp_q\ (s,\ n) \tag{1}$$

Let 'n'', represent the ordered snippets, where one of them has been replaced by another snippet 's' and 'n'' is the perturbation of 'n.' The non-integrative 'NI' and the integrative 'I' of query 'q' can be expressed as a function and is given below:

$$|\ wp_q\ (s,\ n);\ wp_q\ (s,\ n') \tag{2}$$

$$NI_q = fun\ \left(|wp_q\ (s,\ n);\ wp_q\ (s,\ n'),\ n' \in \{s_1,\ s_2,\ \ldots, s_n\}\ \right) \tag{3}$$

$$I_q = g\left(NI_q\right) \tag{4}$$

The iterative procedure '$g()$' involved in the above function is expressed as given below. Let 's' represent a snippet, performed on a query 'q', and let '$s + \varepsilon Q$' represent a new operation that varies slightly from the first, in which 'ε' represents a small threshold constant. If 'q' is a query, then '$sq = Tq$', where 'T' is said to be the threshold constant. The perturbed problem of determining a function 'g' as given below:

$$(s + \varepsilon Q)g = Tg \tag{5}$$

$$(s - T)g = -\varepsilon Qg \tag{6}$$

Then the function 'g_1' that satisfies the equation '$(s - T)g = -\varepsilon Qg$' is called the first approximation to 'g.' The function 'g_2' that satisfies the equation '$(s - T)g_2 = -\varepsilon Qg_2$' is called the second approximation to 'g' and so on, with the 'nth' approximation 'g_n' satisfying '$(s - T)g_n = -\varepsilon Qg_{n-1}$'.

If the sequence '$g_1,\ g_2,\ g_3,\ \ldots., g_n$' converges to a specific function, that function is then said to be the essential solution to the problem. The largest value of 'ε' for which the sequence converges is called the radius of convergence of the solution. Thus, the non-integrative nature of the context-dependent term increases with semantic separation but the integrative nature of context-dependency decreases with semantic separation. The perturbation sets for query 'q' are comprised of snippets '$s_1,\ s_2,\ \ldots s_n$', where 'S_j' is a synonym of 's_j' and is expressed as given below:

$$\{P_1 \ldots . P_n\} = \left\{ \begin{array}{c} S_1 s_2 \ldots . s_n \\ s_1 S_2 \ldots\ s_n \\ . \\ . \\ . \\ s_1 \ldots . s_{j-1}\ S_j\ s_{j+1} \ldots . . s_n \\ . \\ . \\ . \\ s_1 \ldots\ s_{n-1} \ldots . . S_n \end{array} \right\} \tag{7}$$

With the perturbed sets obtained from (7), linear space is generated for semantic data analysis. Let '$\overrightarrow{LS}\ (q)$' and '$\overrightarrow{LS}\ (P_i)$' represent the linear space of query 'q' and its perturbation 'P_i,' respectively. The semantic separation '|' between query 'q' and its perturbation 'P_i' is constructed as the distance between their linear space, with 'Dis' representing the distance containing the snippets '$s_1, s_2, \ldots, s_n$' of 'k' equivalent snippets. Then the function '$fun\ ()$' in (3) of a query 'q' is comprised of snippets '$s_1, s_2, \ldots, s_n$' and is expressed as given below:

$$fun\ () = \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} Dis\left(\overrightarrow{LS}\ (q),\ \overrightarrow{LS}\ (P_{ij})\right) \tag{8}$$

From (8), 'P_{ij}' represents the perturbation '$s_1 \ldots s_{j-1} S_{ij} s_{j+1} \ldots .s_n$' and '$S_{ij}$' is the '$ith$' synonym of snippet '$s_j$' obtained from (7). Once all the snippets are extracted from the query 'q', for corresponding web page 'wp,' the individual snippet influence term and query influence are obtained. The block diagram for Combinatorial Factor (CF) determination is shown in Fig. 2.

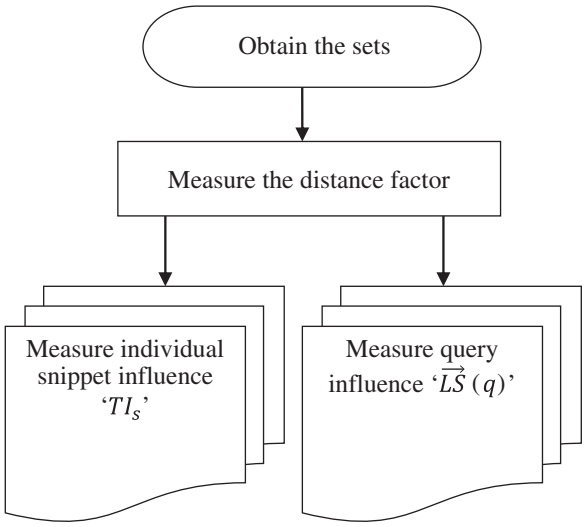

Figure 2: Block diagram of combinatorial factor determination

As shown in Fig. 2, once all the snippets are extracted from the query 'q', for corresponding web-page 'wp' the block diagram of CF determination consists of the perturbation sets obtained during the initial process. With this, the distance factor is computed for extracting equivalent snippets. Finally, the CF individual snippet influence and query influence are obtained. The individual snippet influence in linear space is evaluated and is expressed as given below:

$$TI_s = \log(f_{is}) * \log\left(\frac{n}{wp(s)}\right) \tag{9}$$

From (9), 'TI' represents the term influence of snippet 's' of 'ith' query, 'f' represents the term frequency of snippet 's' of web page 'wp' and 'n' represents the total number of snippets in web page $'wp'$. Having built such a linear space representation for each '$s \in q$', the linear space of the entire query influence in the proposed work is constructed as their point-wise multiplication for effective semantic linear space representation. This is expressed as given below:

$$\overrightarrow{LS}\ (q) = \overrightarrow{LS}\ (s_1)\ \gamma \odot \ldots \odot \overrightarrow{LS}\ (s_n) \tag{10}$$

$$\overrightarrow{LS}\ (q) = (a_1, \ldots, a_b) \odot (b_1, \ldots, b_n) \tag{11}$$

$$\overrightarrow{LS}\ (q) = (x_1 * y_1, \ldots, x_n * y_n) \tag{12}$$

The linear spaces of the snippets on non-integrative queries that will commonly occur in non-identical contexts will have entries with low absolute values. However, for integrative queries, substituting a snippet with its synonym yields constructions that are likely to occur in a number of contexts that are different from the original. They have dissimilar contextual statistics and thus greater distance '***Dis***.' Fig. 3 shows the Linear Space Context Dependent algorithm.

Input: web page ' wp', snippets ' $s_1, s_2, \ldots, s_n$', keyword ' k', number of snippets 'n', query 'q'

Output: detection of strongly dependent snippets from query

```
1: Begin
2:     For each Web Page 'wp'
3:         For each Snippets 'S'
4:             For each query 'q'
5:                 Measure the non-integrative key using (3)
6:                 Measure the integrative key using (4)
7:                 Obtain the perturbation sets for query 'q' using (7)
8:                 Measure the individual snippet influence using (9)
9:                 Measure the query influence using (12)
10:            End for
11:        End for
12:    End for
13: End
```

Figure 3: Algorithm for the linear space context dependent algorithm

The algorithm for strongly identifying the dependent query terms with the aid of non-integrative nature is analyzed and shown in Fig. 4. Individual words in a query do not have a greater influence. However, their meanings differ according to the context. The proposed method uses the non-integrative nature of the query to detect strongly dependent snippets from the given query. Both the influence of snippets and the frequency of snippets are measured to identify strongly dependent snippets with the aid of linear space. This reduces the ambiguity related to inter-sentential context.

3.2 Combinatorial Optimization

With the combinatorial optimized factors, although ambiguity related to inter-sentential context is reduced, the time required to evaluate a query increases. To address this, a knapsack-based combinatorial optimization for semantic data analysis is constructed. Selecting the strongly dependent snippets and inter-sentential context into the cache is a '$0-1$' knapsack problem.

Given a knapsack with capacity 'c' , 'n' items '$c_1, c_2, \ldots, c_n$', having individual snippet influence 'TI_s' and overall query influence '$\overrightarrow{LS}\,(q)$', take the items that maximize the individual snippet influence without exceeding 'c'. A snippet can be selected only if the fractions of items cannot be taken. As the greedy strategy does not always guarantee an optimal solution for the knapsack problem, the proposed work describes how to

formulate the selection of strongly dependent snippets and inter-sentential context as a combinatorial optimization problem. We formulate the knapsack combinatorial optimization problem as an integer programming problem, as given below.

Input: capacity 'c', items 'n', individual snippet influence 'TI_s', overall query influence '$\overrightarrow{LS}(q)$',

Output: Time-optimized semantic data analytics

```
1: Begin
2:     For each Web Page 'wp'
3:         For each Snippets 'S'
4:             For each query 'q'
5:                 Obtain the maximization formulates using (13)
6:                 Design the constraints using (14)
7:             End for
8:         End for
9:     End for
10: End
```

Figure 4: Algorithm for the Knapsack combinatorial optimization algorithm

$$Max \sum_{i=1}^{n} TI_s s_i \tag{13}$$

$$Subject\ to \sum_{i=1}^{n} \overrightarrow{LS}(q) s_i \leq c,\ where\ s_i \in s_1, s_2, .., s_n \tag{14}$$

From (13), '$\sum_{i=1}^{n} TI_s s_i$' is the objective function, '$\overrightarrow{LS}(q) s_i \leq c$' and '$s_i \in s_1, s_2, .., s_n$' are the constraints, where 's_i' represents a snippet. A solution is to set the snippets '$s_{i,}$' a solution that satisfies all the constraints and one that yields maximum objective function value.

The objective behind the design of the proposed work is the consideration of optimal solutions. From (14), the proposed work states that the total snippets cannot exceed the query size or capacity 'c,' whereas (14) states that each snippet is either selected or discarded. Fig. 4 shows the Knapsack Combinatorial Optimization algorithm.

For example, consider a Tripadvisor dataset consisting of reviews randomly selected from several accommodations. In order to obtain the maximization, formulates (13) are used according to the design constraints from (14), with consideration of two snippets: Room file snippets and value file snippets. With these design constraints, optimal solutions are identified, thereby meeting the objectives.

4 Experimental Settings

The queries were simulated and the performance was measured. The COLSK-IR method was evaluated [26] against PolaritySim [6] and DomainOntoWP [13] using the number of reviews as the measurement of our web page performance. We experimented with review sizes of 15, 30, 45, 60, 75, 90, 105, 120, 135 and 150, with 512 bytes of review on the Tripadvisor dataset, which included an overall review of 200 randomly selected accommodations.

The dataset of approximately 200 reviews was taken from Tripadvisor.com through a random selection. It covered all five satisfaction levels (40 reviews in each level) consisting of 1,382 criticisms, 211

non-criticisms and 97 criticisms with errors. The information was collected from Tripadvisor and Edmunds. Tripadvisor had 259,000 reviews.

The experiment was conducted based on factors such as number of reviews, non-integrative key extraction time, recall rate, precision and semantic data analysis efficiency. To evaluate the performance of the COLSK-IR method, two metrics were introduced to measure the semantic data analysis and compared with the existing methods: Polarity Similarity (PolaritySim) and Domain Ontology of Web Pages (DomainOntoWP).

5 Discussion

The performance of COLSK-IR for semantic data analysis was compared with the Polarity Similarity (PolaritySim) and Domain Ontology of Web Pages (DomainOntoWP). The experiments measured the effectiveness of non-integrative key extraction time, precision rate and recall for 150 reviews, using the method described in Section 3.

5.1 Non-Integrative Key Extraction Time

The non-integrative key extraction time measured the time required to extract the non-integrative key (i.e., extracted keys) with respect to the total number of reviews in web pages. The non-integrative key extraction time is measured as given below.

$$NI - KE_t = r_i * Time\left(NI_q\right) \tag{15}$$

From (15), '$NI - KE_t$' refers to the non-integrative key extraction time using the number of reviews 'r_i' for the extracted keywords 'k_i' respectively, measured in terms of milliseconds (ms). Tab. 1 shows the non-integrative key extraction time of the proposed COLSK-IR and the PolaritySim and DomainOntoWP methods. The proposed COLSK-IR method outperformed the existing methods in terms of non-integrative key extraction time.

Fig. 5 shows the results of non-integrative key extraction time vs. the varying number of reviews. To better distinguish the efficacy of the proposed COLSK-IR method, the experimental results are shown in Tab. 1, where it is compared against PolaritySim and DomainOntoWP.

Results are presented for 10 numbers of reviews. The non-integrative key extraction time for these 10 numbers of reviews measures the time taken for convergence on different reviews as in (1). The reported results confirm that with the increase in the number of reviews, the non-integrative key extraction time also increases. The process is repeated for 150 reviews for conducting experiments, as illustrated in Fig. 5. The proposed COLSK-IR method performs relatively well when compared to the PolaritySim and DomainOntoWP methods. The COLSK-IR method offers better changes using its iteration procedure, which considers perturbation as a factor for semantic data analysis by 21% when compared to PolaritySim. Moreover, the approximation factor with perturbation sets in the COLSK-IR method considers both the non-integrative and integrative snippets to reduce the convergent time on semantic data analysis by 45% when compared to DomainOntoWP.

5.2 Precision Rate

Precision rate refers to the number of relevant snippets extracted with respect to the number of returned snippets, i.e.,

$$P = \sum_{i=1}^{n} \left(\frac{Rel(s_i)}{n} \right) * 100 \tag{16}$$

From (16), the precision rate 'P' is obtained, using the relevant snippets extracted, as are '$Rel(s_i)$' and the total number of extracted snippets, 'n' from web pages. Tab. 2 shows the precision rate of the proposed COLSK-IR method for 150 reviews and comparisons made against PolaritySim [2] and DomainOntoWP [13].

Table 1: Non-integrative key extraction time obtained using COLSK-IR, PolaritySim and DomainOntoWP

No. of reviews	Non-integrative key extraction time (ms)		
	COLSK-IR	PolaritySim	DomainOntoWP
15	4.15	7.45	8.3
30	7.13	10.13	12.54
45	11.17	13.14	17.43
60	15.32	17.21	24.24
75	20.13	22.14	30.16
90	28.32	31.32	36.25
105	33.14	35.79	42.39
120	36.14	39.32	43.21
135	38.25	41.15	45.61
150	41.43	44.23	49.12

Table 2: Precision rate obtained using COLSK-IR, PolaritySim, and DomainOntoWP

No. of reviews	Precision rate (%)		
	COLSK-IR	PolaritySim	DomainOntoWP
15	77.51	67.94	61.28
30	74.31	64.30	58.22
45	72.54	62.52	56.44
60	70.38	60.35	54.27
75	68.25	58.22	52.14
90	82.99	72.96	66.87
105	89.95	80.92	74.83
120	90.14	82.14	78.32
135	92.23	87.13	82.14
150	94.14	89.13	86.27

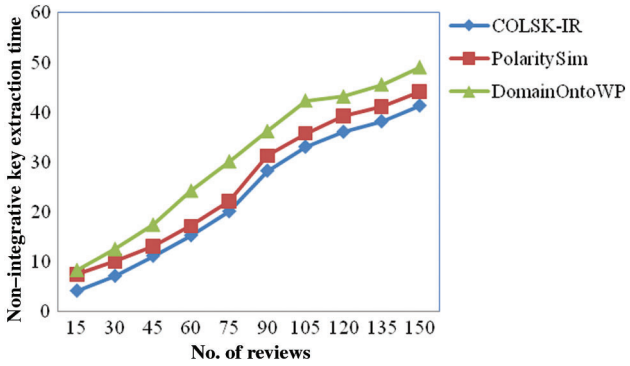

Figure 5: Measure of non-integrative key extraction time

To increase the precision of semantic data analysis for web pages, first approximation, second approximation, and '*nth*' approximation are considered, as shown in Fig. 6. With this, the radius of convergence of the solution that converges to a specific function with approximation factor is included, resulting in the optimal solution according to the number of reviews.

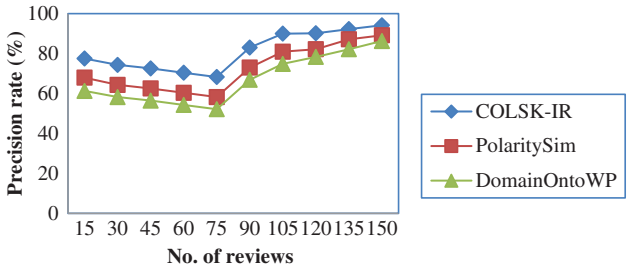

Figure 6: Measure of precision rate

In the experimental setup, the number of reviews ranged from 15 to 150. The results for 10 different types of reviews collected from Tripadvisor and Edmunds are shown in Fig. 7. The precision rate of our COLSK-IR method is comparable to that of the state-of-the-art methods. The precision rate is the ratio of the relevant snippets extracted to the overall snippets considered for semantic data analysis.

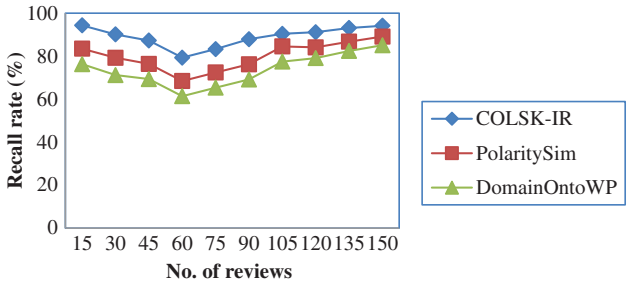

Figure 7: Measure of recall rate

5.3 Recall

Recall rate measures the number of relevant snippets extracted with respect to the number of relevant snippets, i.e., the number of extracted relevant snippets returned by the web page '*wp*' with regard to the '*Rel* (*s*)' returned relevant snippets.

$$R = \sum_{i=1}^{n} \left(\frac{Rel(s_i)}{Rel\ (s)} \right) * 100 \tag{17}$$

From (17), the recall rate 'R' is obtained, using the relevant extracted snippets, '$Rel(s_i)$' and the total number of relevant snippets, 's' from web pages. Here we try to show the examples of reviews yielding the highest and lowest recall rates using the methods COLSK-IR, PolaritySim, and DomainOntoWP methods. Tab. 3 shows the tabulation of recall rates using these three methods.

Table 3: Recall rate obtained using COLSK-IR, PolaritySim, and DomainOntoWP

No. of reviews	Recall rate (%)		
	COLSK-IR	PolaritySim	DomainOntoWP
15	94.36	83.51	76.29
30	90.16	79.29	71.23
45	87.29	76.42	69.36
60	79.33	68.46	61.40
75	83.29	72.42	65.36
90	87.90	76.25	69.19
105	90.43	84.56	77.50
120	91.18	84.13	79.13
135	93.14	86.78	82.45
150	94.18	89.13	85.10

Fig. 7 shows the recall rates for the three methods for reviews increasing in number from 15 to 150. The recall rate improvement of COLSK-IR over PolaritySim and DomainOntoWP decreases gradually as the number of reviews increases, though not linearly. It can be inferred that a further increase is found between the reviews in the range of 75 and 150 because of the presence of noise prior to the key information extraction during semantic data analysis.

As shown in Fig. 7, for example, when the number of reviews is 15, the percentage improvement of COLSK-IR method is 11% compared to PolaritySim is 11% and 19% compared to DomainOntoWP. When the number of reviews is 75, the improvement is around 13% compared to PolaritySim and 22% compared to DomainOntoWP. The reason for this is the application of Combinatorial Factor determination. The knapsack combinatorial optimization problem as an integer programming problem is extended to formulate the selection of strongly dependent snippets and inter-sentential context as a combinatorial optimization problem that extends the recall rate by 17% compared to DomainOntoWP.

6 Conclusion

This paper proposes a Combinatorial Optimized Linear Space Knapsack for Information Retrieval (COLSK-IR) to overcome the difficulty of detecting strongly dependent snippets and reducing the ambiguity related to inter-sentential context. This paper shows how this method can be extended to incorporate the time required to evaluate a query for efficient semantic data analysis based on the knapsack problem. This paper provides two algorithms: Linear Space Context Dependent and Knapsack Combinatorial Optimization. The Linear Space Context Dependent algorithm manages and identifies

strongly dependent snippets based on the influence and frequency of snippets. The Knapsack Combinatorial Optimization algorithm reduces the ambiguity related to inter-sentential context by formulating an integer programming problem to determine the optimal solutions. The experimental results show that the COLSK-IR provides better performance than the state-of-the-art methods in terms of the parameters such as non-integrative key extraction time, precision and recall rate.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Chen, B. Song, L. Fan, X. Du and M. Guizani, "Multi–modal data semantic localization with relationship dependencies for efficient signal processing in EH CRNs," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 2, pp. 347–357, 2019.

[2] O. Vechtomova, "Disambiguating context-dependent polarity of words: An information retrieval approach," *Information Processing & Management*, vol. 53, no. 5, pp. 1062–1079, 2017.

[3] T. T. S. Nguyen, H. Y. Lu and J. Lu, "Web-page recommendation based on web usage and domain knowledge," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2574–2587, 2014.

[4] D. E. Losada, J. Parapar and A. Barreiro, "Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems," *Information Processing & Management*, vol. 53, no. 5, pp. 1005–1025, 2017.

[5] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat and R. Baeza-Yates, "A machine learning approach for result caching in web search engines," *Information Processing & Management*, vol. 53, no. 4, pp. 834–850, 2017.

[6] X. Xie, X. Cai, J. Zhou, N. Cao and Y. Wu, "A semantic-based method for visualizing large image collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 7, pp. 2362–2377, 2019.

[7] D. Song, Y. Luo and J. Heflin, "Linking heterogeneous data in the semantic web using scalable and domain-independent candidate selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 143–156, 2017.

[8] J. Zeleny, R. Burget and J. Zendulka, "Box clustering segmentation: A new method for vision-based web page preprocessing," *Information Processing & Management*, vol. 53, no. 3, pp. 735–750, 2017.

[9] K. Becker, V. P. Moreira and A. G. L. dos Santos , "Multilingual emotion classification using supervised learning: Comparative experiments," *Information Processing & Management*, vol. 53, no. 3, pp. 684–704, 2017.

[10] J. Ziegler and L. Pylkkänen, "Scalar adjectives and the temporal unfolding of semantic composition: An MEG investigation," *Neuropsychologia*, vol. 89, pp. 161–171, 2016.

[11] T. Leffel, M. Lauter, M. Westerlund and L. Pylkkänen, "Restrictive *vs.* non-restrictive composition: A magnetoencephalography study," *Language Cognition and Neuroscience*, vol. 29, no. 10, pp. 1191–1204, 2014.

[12] M. Xiang, J. Grove and A. Giannakidou, "Semantic and pragmatic processes in the comprehension of negation: An event related potential study of negative polarity sensitivity," *Journal of Neurolinguistics*, vol. 38, pp. 71–88, 2016.

[13] M. Nuo, C. Lun and H. Liu, "Tibetan multi-word expressions identification framework based on news corpora," *Natural Language Understanding and Intelligent Applications*, vol. 10102, pp. 16–26, 2016.

[14] J. Li, C. Liu and J. X. Yu, "Context-based diversification for keyword queries over XML data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 660–672, 2015.

[15] W. Zhang, Z. Ming, Y. Zhang, T. Liu and T. Chua, "Capturing the semantics of key phrases using multiple languages for question retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 4, pp. 888–900, 2016.

[16] Z. Dou, Z. Jiang, S. Hu, J. R. Wen and R. Song, "Automatically mining facets for queries from their search results," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 385–397, 2016.

[17] E. Serra and F. Spezzano, "An effective GPU-based approach to probabilistic query confidence Computation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 1, pp. 17–31, 2015.

[18] Y. B. Kang, P. D. Haghigh and F. Burstein, "TaxoFinder: A graph-based approach for taxonomy learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 524–536, 2016.

[19] A. Behnaz, A. Natarajan, F. A. Rabhi and M. Peat, "A semantic-based analytics architecture and its application to commodity pricing, enterprise Applications, markets and Services in the Finance Industry, FinanceCom 2016, " in *Lecture Notes in Business Information Processing*, vol. 276. Cham: Springer, 2016.

[20] T. G. Stavropoulos, E. Kontopoulos, A. M. Peñuela, S. Tachos, S. Andreadis *et al.,* "Cross-domain semantic drift measurement in ontologies using the semadrift tool and metrics," *3rd Workshop on Managing the Evolution and Preservation of the Data Web, CEUR Workshop Proceedings*, Portoroz, Slovenia, 1824, 2017.

[21] A. Borisov, P. Serdyukov and M. de Rijke, "Using metafeatures to increase the effectiveness of latent semantic models in web Search," in *International World Wide Web Conf. Committee, ACM*, Montréal, Québec, Canada, 2016.

[22] L. Bing, Z. Zhang, W. Lam and W. W. Cohen, "Towards a language-independent solution: Knowledge base completion by searching the web and deriving language pattern," *Knowledge-Based Systems*, vol. 115, pp. 80–86, 2017.

[23] A. Borisov, I. Markov, M. de Rijke and P. Serdyukov, "A context-aware time model for web search," in *Proc. of the 39th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Pisa, Italy, pp. 205–214, 2016.

[24] M. Thomas and V. S. Chooralil, "Security and privacy via optimised blockchain," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 3, pp. 415–418, 2019.

[25] V. P. Vijayan, D. John, M. Thomas, N. V. Maliackal and S. S. Vargheese, "Multi agent path planning approach to dynamic free flight environment," *International Journal of Recent Trends in Engineering and Technology*, vol. 1, no. 1, pp. 41–46, 2009.

[26] V. P. Vijayan and B. Paul, "Multi objective traffic prediction using type-2 fuzzy logic and ambient intelligence," *Int. Conf. on Advances in Computer Engineering*, pp. 309–311, 2010.