Data-driven Context Discovery Model for Semantic Computing in the Big Data Era

Takafumi Nakanishi
Center for Global Communications (GLOCOM),
International University of Japan
Tokyo, Japan
e-mail: takafumi@glocom.ac.jp

***Abstract*—We introduce a data-driven context discovery model for semantic computing in the big data era. Our model extracts from data sets the appropriate feature set as the context. We suggest that selection of a target data set is one of the representation processes for this purpose and the context in a big data environment. When a person selects target data from big data, that action latently indicates the context represented by the person's intention. Selecting a feature set from big data constitutes a data-driven context creation. Recently, fragmental data has spread on the Internet. In order to analyze big data, it is necessary to aggregate the appropriate data from data that has been dispersed on the Internet. An aggregation policy represents the purposes or contexts of analysis. In the big data era, it is necessary to focus not only on analysis but also on aggregation. After data aggregation, it is necessary to extract feature sets for semantic computing. This is what our model focuses on.**

Keywords: data-driven; context; feature selection; big data; data set

I. INTRODUCTION

Recently, big data is generated in a number of ways, including Internet browsing, sensors, and smartphones, etc. Most people have said that big data is a big opportunity. However, there are some who get hooked on a flood of big data. We information science researchers have already constructed data sensing, aggregation, retrieval, analysis, and visualization environments via web portals, software, APIs, etc. on the Internet. It is necessary to encourage people to use these big data. The number of information resources available on the Internet has been increasing rapidly. In particular, there is a large amount of fragmental data created by each person's device or created by the number of sophisticated sensors for the sake of scientific curiosity. In short, we not only retrieve but also create these data every day. Mountains of various fragmental data are being created.

One of the most important points is that data has become not only massive but also fragmentary. Currently, most users search contents through a search engine. This means that users acquire pages as contents. As data becomes fragmented, a model that searches for a single page will fail. It is more important to survey the entire data set than to analyze one piece of data deeply, given the large amount of fragmental data.

We observe that the essence of big data is not only massive data processing, but also optimization of the real world through the knowledge acquired from aggregated data. The current tendency of research on big data is how to aggregate a massive amount of data and how to process the data quickly. In the future, research will tend to focus on methods of discovering optimized solutions from big data.

Meanings are relatively determined by the context in a dynamic manner. One of the most important issues is achieving dynamic semantic computing that depends on the context. The dynamic nature is a very important part of the essence, because data that represents the features of each concept changes on each occasion and in each location. In other words, big data has volume, velocity, and variability. In order to compute semantics, a process to determine a context as a viewpoint is required. This means that it is necessary to predefine a space for the measurement of correlation. The space consists of feature sets as axes. Because we cannot predefine the feature set, it is necessary to develop a method of data-driven feature selection for semantic computing. The selected feature set constructs a measurement space. In other words, the measurement space represents the context in semantic computing.

In this paper, we introduce a data-driven context discovery model for semantic computing in the big data era. Our model extracts from data sets the appropriate feature set as the context.

We suggest that selection of a target data set is one of the representation processes for this purpose and the context in a big data environment. When a person selects target data from big data, that action latently indicates the context represented by the person's intention. Selecting a feature set from big data constitutes a data-driven context creation. Recently, fragmental data has spread on the Internet. In order to analyze big data, it is necessary to aggregate the appropriate data from data that has been dispersed on the Internet. An aggregation policy represents the purposes or contexts of analysis. In the big data era, it is necessary to focus not only on analysis but also on aggregation. After data aggregation, it is necessary to extract feature sets for semantic computing. This is what our model focuses on.

The contributions of our paper are as follow.

- We propose a new model of semantic computing by achieving a data-driven feature selection.
- The system applied to our model extracts feature sets corresponding to data sets, because an aggregation policy represents purposes or contexts of analysis.

- Our method reduces the computational cost of measurement of semantic computing because our method reduces the dimension of each vector represented in a certain selected feature set.

This paper is organized as follows. In Section II, we survey the existing work related to our proposed method. In Section II, we present the basic idea of our model. Next, we describe formulation of the design of our model in Section IV. Finally, we present our conclusions in Section V.

II. Related Work

One of the most important issues of semantic computing is correlation and similarity measurement. The most popular and basic method is utilization of the vector space model [1]. The dimensionality reduction techniques of the vector space model have been used for developing traditional vector space models, such as latent semantic indexing [2].

A weighting method is regarded as one of the feature selection techniques. Reference [3] describes a survey of weighting methods, such as binary [4], term frequency (TF) [4], augmented normalized term frequency [4][5], log [5], inverse document frequency (IDF) [4], probabilistic inverse [4][5], and document length normalization [4].

There have been studies defining similarity metrics for hierarchical structures such as WordNet [6]. Rada et al [7] have proposed a “conceptual distance” that indicates the similarity between concepts of semantic nets by using path lengths. Some studies [8][9] have extended and used the conceptual distance for information retrieval. Resnik [10] has proposed an alternative similarity measurement based on the concept of information content. Ganesan et al [11] have presented new similarity measurements in order to obtain similarity scores that are more intuitive than those based on traditional measurements.

In regard to other perspectives, the reference [12] has been surveyed. This survey [12] shows common architecture and general functionality as OBIE from various ontology-based information extraction studies. It consists of an “information extraction module,” “ontology generator,” “ontology editor,” “semantic lexicon,” and a number of preprocessors. The researchers are working both on various studies of OBIE system implementation and on studies focused on each module. In this paper, we will mainly introduce research on OBIE system implementation.

Our model processes a dynamic data-driven feature selection corresponding to a context. This means that our model does not have to prepare the space in advance. This is a very important difference, because we cannot create the space or schemas in advance in an open assumption. Currently, we are in the big data era. In a big data environment, we can aggregate a large amount of diverse fragmental data. We cannot predict in advance the kinds of data we will obtain. In fact, an increased key-value store means that the schema cannot be designed in advance. Since data updates are increasing in speed, the space for semantic computations and analyses should change dynamically as well.

One of the more famous methods of feature selection is “bags of keypoints” [13]. The bag of keypoints method is based on vector quantization of affine invariant descriptors of image patches. We can use the bag of keypoints for image classification.

An overview of feature selection algorithms is given in reference [14]. In this case, the feature selection algorithm is a computational solution that is motivated by a certain definition of relevance. It is hard to define the relevance. This [14] represents some roles of feature selection as follows: 1) Search organization, 2) Generation of successors, and 3) Evaluation measure.

Type 1) is in relation to the portion of the hypothesis explored with respect to their total number. This is responsible for driving the feature selection process using a specific strategy. The methods related to type 1) are [15], [16], and [17]. Type 2) proposes possible variants (successor candidates) of the current hypothesis. The method related to type 2) is [18]. Type 3) compares different hypotheses to guide the search process. The methods related to type 3) are [19], [20], and [21].

[14] also represents a general scheme for feature selection. The relationship between a feature selection algorithm and the inducer chosen to evaluate the usefulness of the feature selection process can take three main forms: embedded, filter, and wrapper.

There are some methods without feature selection, such as deep learning [22]. However, it is not possible to ignore feature selection completely. Generally, an artificial intelligence must depend on evaluation functions that are created by humans. The evaluation function is dependent on the manner in which features are selected. Even if more work is done on deep learning, work related to feature selection will still be conducted.

Currently, we are in the big data era. In a big data environment, we can aggregate a large amount of diverse fragmental data. We cannot predict in advance the kinds of data we will obtain. In fact, an increased key-value store means that the schema cannot be designed in advance. Since data updates are increasing in speed, the space for semantic computations and analyses should change dynamically as well.

Our model clearly differs in purpose from other methods. The current method predefines semantics as a measurement space, ontology, etc. By contrast, the system applied in our method extracts an appropriate feature set from a given data set. The given data set is the target data set. We suggest that selection of a target data set is one of the representation processes for this purpose and the context in a big data environment. Meanings are relatively determined by the context in a dynamic manner. One of the most important issues is achieving dynamic semantic computing that depends on the context. The dynamic nature is a very important part of the essence, because data that represents the features of each concept changes on each occasion and in each location.

Selecting a feature set from big data constitutes a data-driven context creation. Recently, fragmental data has spread on the Internet. In order to analyze big data, it is necessary to aggregate the appropriate data from data that has been dispursed on the Internet. An aggregation policy represents the purposes or contexts of analysis. In the big data era, it is necessary to focus not only on analysis but also o aggregation. After data aggregation, it is necessary to extract feature sets for semantic computing. This is what our model focuses on.

We have proposed a new weighting method for the vector space model [23]. This paper presents an overview of the reference [23]. In particular, the system that has been applied to our model extracts feature sets corresponding to data sets, because an aggregation policy represents purposes or contexts of analysis.

III. Basic Idea of Our Model

In this section, we introduce our assumptions and basic ideas for our model: a data-driven context discovery model for semantic computing.

Figure 1 shows an overview of the basic idea for our model.

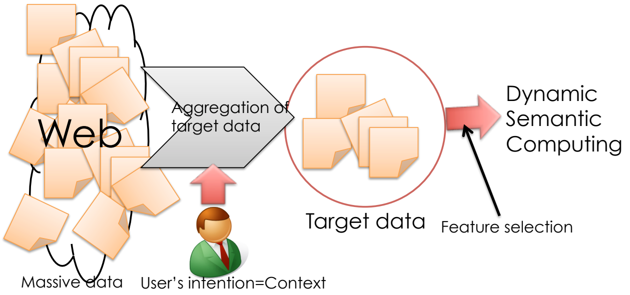

Figure 1. Basic idea of our model.
There is a large amount of data on the Internet. When we would like to analyze something, we try to aggregate data. In this case, we suggest that selection of a target data set is one of the representation processes for this purpose and the context in a big data environment. When a person selects target data from big data, that action latently indicates the context represented by the person's intention. In other words, the system can dynamically extract semantics by extracting feature sets. We can analyze dynamic data-driven semantic computing.

The dynamic nature is a very important part of the essence, because data that represents the features of each concept changes on each occasion and in each location. In other words, big data has volume, velocity, and variability. In order to compute semantics, a process to determine a context as a viewpoint is required. This means that it is necessary to predefine a space for the measurement of correlation. The space consists of feature sets as axes. Because we cannot predefine the feature set, it is necessary to develop a method of data-driven feature selection for semantic computing.

We suggest that selection of a target data set is one of the representation processes for this purpose and the context in a big data environment. When a person selects target data from big data, that action latently indicates the context represented by the person's intention.

Recently, fragmental data has spread on the Internet. In order to analysis big data, it is necessary to aggregate the appropriate data from data that has been dispersed on the Internet. In this case, we use crawler techniques. More specifically, we use focused crawlers. The focused crawler aggregates data corresponding to conditions given by a user. Therefore, this process includes the user's intention. The user's intention is one of the important clues for context detection.

The system applied to our model detects context from aggregated data because of this background. Context detection is achieved through feature selection.

We suggest that feature sets create the context. The feature set can construct measurement space. Each feature is driven by each axis of the measurement space. The measurement space achieves similarity or correlation of semantics. For example, the system detects the context of correlation between climate and another factor when we aggregate temperature data. Therefore, we can identify the context through aggregated data.

In other words, we can extract semantics from data usage logs. Figure 2 shows the relationship between content, context, and semantics.

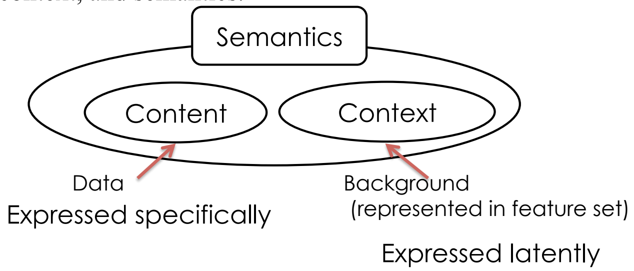

Figure 2. Relationship between semantics, content, and context. Semantics consist of content and context. Content is something expressed specifically, such as data itself. Context is something expressed latently. The system applied to our model extracts feature sets from data sets. In other words, we can identify the context through data set usage. Semantics are created by data itself and data usage in our model.

Semantics consist of content and context. Content is something expressed specifically, such as data itself. Context is something expressed latently. The system applied to our model extracts feature sets from data sets. In other words, we can identify the context through data set usage. Semantics are created by data itself and data usage in our model.

Data usage logs represent context. It is important to achieve dynamic semantic computing. Semantics consist of content and context. We can aggregate data on the Internet as content. The system applied to our method can extract feature sets as context. Therefore, we can identify semantics by content and context.

Please note that the semantics of data change dynamically through usage of the same data. In this model, data is content. The same data has various ways in which it can be used. When the method of use changes, the context also changes. Therefore, the semantics of data change dynamically.

This is an importance element of dynamic semantic computing. Semantics are relatively determined by the context in a dynamic manner. One of the most important issues is achieving dynamic semantic computing that depends on the context. The dynamic nature is a very important part of the essence, because data that represents the features of each concept changes on each occasion and in each location.

It is difficult to extract context. This paper addresses how to extract context. The system applied to our model is one solution for extracting context. When a person would like to analyze something, he or she selects target data from big data.

IV. Formulation of the Data-driven Context Discovery Model

In this section, we present our method: a data-driven context discovery model for semantic computing in the big data era. Our model extracts the appropriate feature sets as the context from data set. We suggest that selection of a target data set is one of the representation processes for this purpose and the context in a big data environment. When a person selects target data from big data, that action latently indicates the context represented by the person's intention. Selecting a feature set from big data constitutes a data-driven context creation.

First, we introduce Bayesian variance, which is used in our model, in Section IV-A. Next, we design a mathematical formulation in Section IV-B.

A. Variational Bayesian Estimation

In this section, we show one of the estimation methods [24]: variational Bayesian estimation, which is used in this paper. Please note that our method can be applied to other estimation methods. In this paper, we use this as the estimation method for a conditional probability set of $p(v_l|e_m)$.

It is expressed with the stochastic variables X and Z. In addition, X is a known stochastic variable and Z is an unknown variable. The unknown variable Z denotes marginalization as follows.

$$p(X) = \int_Z p(X,Z)dZ$$

$$logp(X) = \int_Z logp(X,Z)\,dZ = L(q) + KL(q||p) \geq L(q)$$

$$L(q) = \int_Z q(Z)\frac{p(X,Z)}{q(Z)}dZ$$

$$KL(q||p) = -\int_Z q(Z)\frac{p(Z|X)}{q(Z)}dZ$$

$KL(q\|p)$ is a Kullback–Leibler divergence. Therefore, the Kullback–Leibler divergence is a minimum value when $q(Z)$=$p(Z|X)$. However, it is difficult to solve $p(Z|X)$ distribution.

Here, we apply it to the mean field approximation. The mean field approximation is represented as follows when a set of unknown variable Z={ z_1,z_2,...,z_k }:

$$q'(Z) = \prod_{i=1}^{k} q_i(z_i)$$

$q'(Z)$ can be represented by the Kullback–Leibler divergence. The approximate solution which should be calculated is equivalent to the minimum of the following formula.

$$KL(q'||q) = \int q'(Z)log\frac{q'(Z)}{q(Z)}dZ$$

he $q'(Z)$ is substituted for $L(q)$:

$$L(q) = \int_Z \prod_{i=1}^{k} q_i(z_i)\frac{p(X,Z)}{\prod_{i=1}^{k} q_i(z_i)}dZ$$

$$= \int_Z q_j(z_j)\left\{\int logp(X,Z)\prod_{i\neq j} q_i(z_i)dz_i\right\}dz_j$$

$$-\int_Z q_j(z_j)log\,q_j(z_j)dz_j + const$$

$$= \int_Z q_j(z_j)log\frac{\tilde{p}(X,z_j)}{q_j(z_j)}dz_j + const = KL(q_j||\tilde{p}) + const$$

Maximization of L(q) is equivalent to minimization of Kullback-Leibler divergence. The optimal solution $q_j^*(Z_j)$ is calculated as follows.

$$logq_j^*(Z_j) = \int logp(X,Z)\prod_{i\neq j} q_i(Z_i)dZ_i + const = \mathbb{E}_{i\neq j}[logp(X,Z)] + const$$

$$q_j^*(Z_j) = \frac{exp(\mathbb{E}_{i\neq j}[logp(X,Z)])}{\int exp(\mathbb{E}_{i\neq j}[logp(X,Z)])dZ_j}$$

B. Formulation of our model

1) Overview

Figure 3 shows an overview of our model. Figure 3 represents the processes in each step.

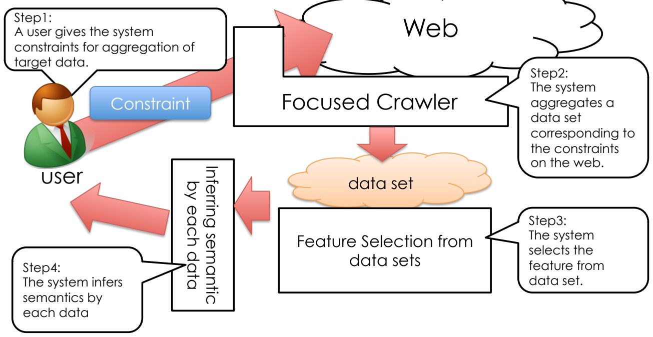
Step1:
A user gives the system constraints for aggregation of target data.
Web
Constraint
Focused Crawler
Step2:
The system aggregates a data set corresponding to the constraints on the web.
user
data set
Inferring semantic by each data
Feature Selection from data sets
Step3:
The system selects the feature from data set.
Step4:
The system infers semantics by each data

Figure 3. Overview of our model.
Our model consists of four steps. Step 1 is giving constraints for data aggregation. Step 2 is aggregation of data corresponding to the constraints on the Internet. Step 3 is feature selection from an aggregated data set. Step 4 is inferring the semantics of each piece of data. In other words, we can add semantics tags for each piece of data

Our model consists of four steps. These steps are as follows.

- Step 1: A user gives the system constraints for aggregation of target data.
  The important point of our model is the utilization of the usage data log. We suggest that feature sets create the context. The feature set can construct measurement space. Each feature is driven by each axis of the measurement space. The measurement space achieves similarity or correlation of semantics. Therefore, first, the user gives the system constraints for the focused crawler. This means that the user defines the usage data.
- Step 2: The system aggregates a data set corresponding to the constraints on the Internet.
  The system aggregates a data set along with the given constraint. The data set represents context. Each piece

of data represents content. When we combine them, we can obtain the semantics of each piece of data.

- Step 3: The system selects features from the data set.
  The system processes the feature selection from the aggregated data set. This step extracts feature sets as semantics axes. The feature set creates the context. We can map each piece of data into a space that is created by the feature set as each axis.
- Step 4: The system infers the semantics of each piece of data.
  We obtain each context and content from the data set and each piece of data. By combining these; we can obtain the semantics of each piece of data by probabilistic weighting.

2) *Formulation*

In this section, we formulate the model in accordance with each step. Figure 4 shows a representation of a graphical model for our model.

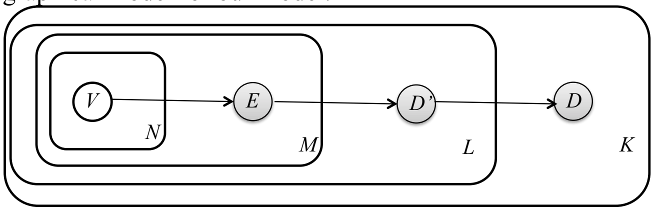

Figure 4. A graphical model of our model
D is a data set on the Internet. *D'* is an aggregated data set corresponding to the given constraint. *E* is an element set of the aggregated data set. *V* is a feature set.

We define the entire data set on the Internet $D=\{d_g\}$, the aggregate data set $D'=\{d'_h\}$, the element set $E=\{e_i\}$, and the feature set $V=\{v_j\}$. We reason $V=\{v_j\}$, when we aggregate D'.

Each element between $\{ d_g \}$ and $\{ d'_h \}$ which is represented in nodes is connected by the edges. Each value of each edge is represented in $p(d_g \mid d_h)$. The number of hit pages of a search engine can predict the values.

Each element between $\{ d'_h \}$ and $\{ e_m \}$ which is represented in nodes is connected by the edges. Each value of each edge is represented in $p(d_h| e_m)$. In the case of a text data set, it is easy to solve. For example, each piece of data has a word. The words are regarded as elements. This means that these values are solved by counting word frequency.

Each element between $\{ v_l \}$ and $\{ e_m \}$ which is represented in nodes is connected by the edges. Each value of each edge is represented in $p(v_l \mid e_i)$.

In conclusion, a conditional probability set of $p(v_l| e_m)$ is a good estimation function of feature selection. In other words, when a conditional probability set of $p(v_l| e_m)$ is bigger than the threshold, we can regard e_m as an appropriate feature v_l.

It is necessary to solve a conditional probability set of $p(v_l| e_m)$. A number of estimation methods in machine learning, such as variational Bayesian estimation [24], etc., are shown in Section III-C. In this paper, we use variational Bayesian estimation [24] as the estimation method for a conditional probability set of $p(v_l| e_m)$.

Finally, we can drive $p(v_l| e_m, d'_h, d_g)$ with the above values. These are represented by the data's metadata. In other words, we can add context-dependent semantics for each piece of data.

V. CONCLUSION

In this paper, we presented a data-driven context discovery model for semantic computing in the big data era. Our model extracts from data sets the appropriate feature set as the context.

We suggest that selection of a target data set is one of the representation processes for this purpose and the context in a big data environment. When a person selects target data from big data, that action latently indicates the context represented by the person's intention.

Our model clearly differs in purpose from other methods. The current method predefines semantics as a measurement space, ontology, etc. By contrast, the system applied to our method extracts an appropriate feature set from a given data set. The given data set is the target data set. We suggest that selection of a target data set is one of the representation processes for this purpose and the context in a big data environment. Meanings are relatively determined by the context in a dynamic manner. One of the most important issues is achieving dynamic semantic computing that depends on the context. The dynamic nature is a very important part of the essence, because data that represents the features of each concept changes on each occasion and in each location.

In the near future, our model will be applied to a heterogeneous data environment. It is necessary to consider the actual application of our model. Dynamic and automatic feature selection, such as is part of our model, is a more important technology in the big data era.

REFERENCES

[1] G. Salton, A. Wong, C. S. Yang, "A vector space model for automatic indexing," *Magazine Communications of the ACM CACM* Homepage archive, vol.18(11), pp. 613-620, Nov. 1975.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, vol. 41(6), pp. 391-407, 1990.

[3] T. G. Kolda, D. P. O'Leary, "A semidiscrete matrix decomposition for latent semantic indexing information retrieval", *Journal ACM Transactions on Information Systems (TOIS)* TOIS Homepage archive vol.16(4), pp. 322-346, Oct. 1998.

[4] G.Salton, C. Buckley, "Term weighting approaches in automatic text retrieval," *Inf. Process. Manage*. 24, pp. 513–523, 1988.

[5] D. Harman, "Ranking algorithms. In Information Retrieval: Data Structures and Algorithms," *W. B. Frakes and R. Baeza-Yates, Eds. Prentice Hall, Englewood Cliffs*, NJ, pp. 363–392, 1992.

[6] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller. "Introduction to WordNet: An on-line lexical database," *Journal of Lexicography*, vol.3(4), pp. 235-244, January 1990.

[7] R. Rada, H. Mili, E. Bicknell, M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man and Cybernetics*, vol.19(1), pp. 17-30, Jan/Feb 1989.

[8] Y. Kim, J. Kim, "A model of knowledge based information retrieval with hierarchical concept graph," *Journal of Documentation*, vol.46(2), pp. 113-136, 1990.

[9] J. Lee, M. Kim, Y. Lee, “Information retrieval based on conceptual distance in is-a hierarchies,” *Journal of Documentation*, vol.49(2), pp. 188-207, 1993.

[10] P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy” In IJCAI: *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 448-453, 1995.

[11] P. Ganesan, H. Garcia-Molina, J. Widom, “Exploiting hierarchical domain structure to compute similarity,” *ACM Trans. Inf. Syst.*, vol.21(1), pp. 64-93, 2003.

[12] D. Wimalasuriya, D. Dou, “Ontology-based information extraction: An introduction and a survey of current approaches,” *Journal of Information Science*. 36, 3 (June 2010), pp. 306-323, 2010.

[13] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, "Visual Categorization with Bags of Keypoints," In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, 2004.

[14] L.C. Molina, L. Belanche, A. Nebot, "Feature selection algorithms: a survey and experimental evaluation," In *Proceedings*. 2002 IEEE International Conference on Data Mining,(ICDM 2003), pp. 306--313, 2002..

[15] P. Narendra and K. Fukunaga, “A Branch and Bound Algorithm for Feature Subset Selection,” *IEEE Transactions on Computer*. C-26(9):917-922, 1977.

[16] J. Pearl. Heuristics, Addison Wesley, 1983.

[17] H. Liu and H. Motoda, “Feature Selection for Knowledge Discoverv nnd Dam Mining. Kluwer Academic Publishers,” I London. GB, 1998.

[18] D. Koller and M. Sahami, “Toward Optimal Feature Selection,” In *Proceedings of the 13th International Conference on Machine Learning*, pp. 284-292, Bari, IT. 1996.

[19] P.A. Devijver and J. Kittler, “PonernRecognition-A Statistical Appmach,” Prentice Hall, London. GB, 1982.

[20] H. Almuallim and T. G. Dietterich, “Leaming Boolean Concepts in the Presence of Many Irrelevant Features,” Artificial Intelligence, 69(1-2):279-305. 1994.

[21] H. Liu and H. Motoda, “Feature Selection for Knowledge Discovery and Dam Mining,” Kluwer Academic Publishers, London. GB, 1998.

[22] Q. V. Le, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, A. Y. Ng, “Building High-level Features Using Large Scale Unsupervised Learning,” In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[23] T. Nakanishi, “Semantic Context-Dependent Weighting for Vector Space Model,” In *Proceedings of the 2014 IEEE International Conference on Semantic Computing (ICSC)*, pp. 262-266, 2014.

[24] M. Christopher Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.