Semantic Similarity Measures in the Biomedical Domain by Leveraging a Web Search Engine

Sheau-Ling Hsieh, Wen-Yung Chang, Chi-Huang Chen, and Yung-Ching Weng

***Abstract*—Various researches in web related semantic similarity measures have been deployed. However, measuring semantic similarity between two terms remains a challenging task. The traditional ontology-based methodologies have a limitation that both concepts must be resided in the same ontology tree(s). Unfortunately, in practice, the assumption is not always applicable. On the other hand, if the corpus is sufficiently adequate, the corpus-based methodologies can overcome the limitation. Now, the web is a continuous and enormous growth corpus. Therefore, a method of estimating semantic similarity is proposed via exploiting the page counts of two biomedical concepts returned by Google AJAX web search engine. The features are extracted as the co-occurrence patterns of two given terms P and Q, by querying P, Q, as well as P AND Q, and the web search hit counts of the defined lexico-syntactic patterns. These similarity scores of different patterns are evaluated, by adapting support vector machines for classification, to leverage the robustness of semantic similarity measures. Experimental results validating against two datasets: dataset 1 provided by A. Hliaoutakis; dataset 2 provided by T. Pedersen, are presented and discussed. In dataset 1, the proposed approach achieves the best correlation coefficient (0.802) under SNOMED-CT. In dataset 2, the proposed method obtains the best correlation coefficient (SNOMED-CT: 0.705; MeSH: 0.723) with physician scores comparing with measures of other methods. However, the correlation coefficients (SNOMED-CT: 0.496; MeSH: 0.539) with coder scores received opposite outcomes. In conclusion, the semantic similarity findings of the proposed method are close to those of physicians' ratings. Furthermore, the study provides a cornerstone investigation for extracting fully relevant information from digitizing, free-text medical records in the National Taiwan University Hospital database.**

***Index Terms*—Semantic similarity, support vector machine, page-count-based, corpus-based, web search engine.**

I. Introduction

NATIONAL Taiwan University Hospital (NTUH), a prestigious teaching hospital in Taiwan, was established in 1895. Currently, on average, there are approximately 9700 outpatients, 300 emergency cases, and 2500 inpatients daily. The NTUH Hospital Information System (HIS) plays an essential role to provide the hospital daily routines and the clinical education. In order to cope with the advancing technologies and operational demands, the centralized HIS has been under transforming since year 2004; the newly distributed HIS has been developed, deployed, and entering fully operations in February 2009. The system aggregates and integrates the entire hospital's functionalities among all departments. For example, it embraces user friendly browser accessibilities, web-based applications focusing on patient cares, pharmacy, laboratory, radiology, administrative activities, financial or billing services, as well as facilities or resource management.

After the popularity of utilizing the NTUH HIS, many electronic medical records are accumulated in the database. In the records, they enclosed physicians' treatment experience and expertise. To raise and enhance the medical quality, reduce costs, as well as support further researches upon these data, a clinical decision support system is established in NTUH to effectively and efficiently extract the knowledge. For example, a liver cancer staging system has been constructed according to the AJCC (American Joint Committee on Cancer) Cancer Staging Manual. In the system, based upon the semantic-driven keyword matching, it can retrieve the number of tumors and their sizes; the data were queried and searched across the radiology reports, operation notes, or discharge summaries. On the surface, it seems easy to apply further research on these electronic records. Unfortunately, the records are stored and remained in free-text formats without following any formal conventions.

On the other hand, previously there are techniques [1], [2] proposed to mining free-text medical reports. However, lots semantically similar terms exist in these reports. For example, "hepatocellular carcinoma" may be written as "hepatoma," "HCC," or "liver cancer." The generic text mining techniques or queries cannot regard these terms as identical ones or synonyms. By simply replacing them, it can either ignore valuable information or interfere with the mining processes among the data [3]. To extract fully relevant information from these electronic, free-text medical reports for further research, a natural language processing technique to measure the semantic similarity under the biomedical domain is required.

II. Related Work

Semantic similarity refers to human judgments of the degrees of relatedness between a given pairs of concepts [3]. There are two main categories: ontology-based and corpus-based. The first class of the techniques is to measure the semantic similarity of the two concepts by calculating the distance between the concept nodes in an ontology tree or hierarchy [4], [5].

Manuscript received May 5, 2012; revised August 27, 2012 and December 27, 2012; accepted March 11, 2013. Date of publication April 25, 2013; date of current version June 27, 2013.

S.-L. Hsieh is with National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: sl_hsieh@cc.nctu.edu.tw).

W.-Y. Chang was with the College of Computer Science, National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail:changuniong@yahoo.com.tw).

C.-H. Chen was with the Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: vinchen@ntu.edu.tw).

Y.-C. Weng is with the Information Office, National Taiwan University Hospital, Taipei 100, Taiwan (e-mail: kimweng59@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2013.2257815

For examples, initially, Rada *et al.* [6] utilized MeSH as semantic networks and devised a "semantic distance" organized in a hierarchy applying to biomedical domains. Lord *et al.* [7] investigated the WordNet-based measures across the gene ontology and discovered that proteins have a high correlation with "sequence similarity." Furthermore, Al-Mubaid and Nguyen [8] explored and presented a strategy to measure the similarity of terms dispersed among multiple ontology. The ontology-based techniques have a limitation that both concepts must be resided in the same tree(s).

The second class of the techniques is to measure the information content of the two concepts as a function representing the probabilities of their occurrences in a corpus. The approach adapts machine learning, rule-based, statistical-based, or web information–based methodologies to analyze the information available and to measure their semantic similarity [3], [9]. By deploying a corpus-based approach, the size of corpus is an essential issue as indicated in many Natural Language Processing (NLP) tasks. Banko and Brill [10] explored a very large corpus as an alternative of implementing sophisticated algorithms. They demonstrated an approach on a lexical disambiguation problem. The problem was to choose which of 2-3 commonly confused concepts were appropriate for a given context. It illustrated that by applying a very simple algorithm, the results continued improving log-linearly with additional training data, even reaching a billion concepts. Thus, they conclude that getting more data may be a better approach than fine tuning the algorithms.

The web is providing unprecedented accessibility to the information as well as interacting with people's daily lives. Today, the obvious source of largest data is the web. Adapting the web as training and testing corpus has gained increasing interests in recent years [9], [11], [12]. Moreover, the web has been applied as a corpus for a variety of NLP tasks, e.g., extraction of semantic relations [13]. Cao and Li [11] proposed a method solving the base noun phrase translation problem by searching translation candidates from the web. Chklovski and Pantel [12] mined the web pages to gather the relations between the semantic verbs. Bollegala *et al.* [9] proposed a methodology to measure the similarity between words that exploit page counts and text snippets returned by a web search engine.

Furthermore, Sánchez *et al.* [14] proposed and evaluated the usage of the web as a background corpus for measuring the similarity of biomedical concepts. Their experiment results indicated that the similarity values obtained from the web are more reliable than those obtained from specific clinical data, manifesting the suitability of the web as an information corpus for the biomedical domain.

However, similarity only considers subsumption relations to assess how two objects are alike, relatedness takes into account a broader range of relations (e.g., part-of). Pirró and Euzenat [15] focused on computing similarity and relatedness by exploiting features of object concepts, expressed in terms of information content and their positions in an ontology structure. By taking into account relations beyond inheritance, they devised a framework that enables to rewrite existing similarity measures that can be augmented to compute semantic relatedness.

Measures of semantic similarity are techniques that attempt to imitate human judgments. In addition, the web obviously can be the candidate of a sufficiently adequate corpus. Therefore, the research adapts a corpus-based approach, utilizing the Google indexed pages as the training corpus. The semantic similarity measures are explored, by five co-occurrence and ten lexico-syntactic patterns, based upon page hit counts of two concepts via web search engine, as input features. The approach applies support vector machines to leverage the robustness of semantic similarity measures.

In the following sections of the paper, initially, the technological background of the study is introduced. The design and methodologies are illustrated in Section IV. In Section V, the experiments are performed and described. In Section VI, the results are presented. Discussions and comparisons with other approaches are elaborated in Section VII. Finally, the paper concludes in Section VIII.

III. Background

In the study, the Google AJAX search API modules are adapted to retrieve the page counts of a given term. Applying Google search engine, the wget() program is used by entering a query term as the input parameter. The search engine returns a file embedded with the page count of the given term; by parsing the returned file, the value of the page count can be obtained. The support vector machines (SVMs) [16], [17] are the machine learning algorithms in the study by implementing different kernel functions. The medical datasets as experimental classifications by SNOMED-CT [18] and MeSH [19] are validated.

IV. Design and Methodology

In the paper, a methodology for semantic similarity measures between two biomedical concepts via page hit counts is proposed by applying the Google web search engine. The designed methodology is described as follows. In Section IV-A, it illustrates the construction of the training sets for classifications. In Section IV-B, the definitions of the input features of the classifiers are explained. The feature selection strategy is elaborated in Section IV-C. The features are ranked by F-score [16] according to their abilities to express semantic similarities. Two-class support vector machines are presented in Section IV-D.

A. Data Preparation and Collection

In order to train the classifiers, both synonymous and nonsynonymous training sets are required. Two websites provide training sets for classifications. At the beginning, the medical terms of the training sets are obtained from the online MedTerms Dictionary of the MedicineNet.com website [20].

To construct the synonymous training sets, a term from MedTerms Dictionary is selected randomly; afterward, the synonyms of the term are searched by querying the Synonyms.net website [21]. The procedure iterates until 1500 synonymous term pairs are accumulated. For collecting the nonsynonymous training set, two sample instances are randomly selected from

MedTerms Dictionary. Both instances are validated via Synonyms.net [21] to ensure that the term pair is not synonymous. The same process iterates until 1500 nonsynonymous term pairs are accumulated and constructed.

MedTerms medical dictionary is the medical reference from MedicineNet.com. Doctors define difficult medical language in easy-to-understand explanations. The training data selected from this website have been reviewed by doctors, and the scope of the datasets has no specific coverage.

B. Feature Definitions

In the study, 15 input features are defined for classifications, including five co-occurrence-based and ten lexico-syntactic pattern-based features. To estimate the co-occurrence of the two concepts P and Q on the web, the corresponding search engine's page counts are collected and calculated. Five popular modified co-occurrence measures are selected [9]: Dice, Jaccard, Overlap (Simpson), point-wise mutual information (PMI), and normalized Google distance (NGD) to compute semantic similarity of the two concepts. The notation $H(P)$ denotes the page counts of query P. The equations of the five co-occurrence-based features are explained as follows:

WebDice coefficient is a variant of the Dice coefficient. WebDice(P, Q) is defined as

$$\text{WebDice}(P,Q) = \begin{cases} 0, & \text{if } H(P \cap Q) \leq c \\ \dfrac{2H(P \cap Q)}{H(P) + H(Q)}, & \text{otherwise.} \end{cases} \tag{1}$$

The WebJaccard coefficient between concepts P and Q, WebJaccard(P, Q), is defined as

$$\text{WebJaccard}(P,Q) = \begin{cases} 0, & \text{if } H(P \cap Q) \leq c \\ \dfrac{H(P \cap Q)}{H(P) + H(Q) - H(P \cap Q)}, & \text{otherwise.} \end{cases} \tag{2}$$

WebOverlap and WebOverlap(P, Q) are defined as

$$\text{WebOverlap}(P,Q) = \begin{cases} 0, & \text{if } H(P \cap Q) \leq c \\ \dfrac{H(P \cap Q)}{\min(H(P), H(Q))}, & \text{otherwise.} \end{cases} \tag{3}$$

WebOverlap is the modification of the Overlap (Simpson) coefficient.

WebPMI is defined as a form of PMI using page counts as follows:

$$\text{WebPMI}(P,Q) = \begin{cases} 0, & \text{if } H(P \cap Q) \leq c \\ \log_2\left(\dfrac{\frac{H(P\cap Q)}{N}}{\frac{H(P)}{N}\frac{H(Q)}{N}}\right), & \text{otherwise.} \end{cases} \tag{4}$$

Here, N is the number of documents indexed by Google. Probabilities in (4) are estimated by the maximum likelihood principle. To calculate PMI accurately by (4), the N value, i.e., the number of documents indexed by Google, is required.

The normalized Google distance (NGD) [22] between concepts P and Q, NGD(P, Q), is defined as (5) shown at the bottom of the page.

Here, N is the number of documents indexed by Google.

McCrae and Collier [23] proposed a method that automatically generates regular expression patterns. It expands seed patterns in a heuristic search and then develops a feature vector depending on the occurrence of pairs in each pattern. Eleven patterns have been devised in [23]; by replacing $*$ with the empty string, the research defined ten lexico-syntactic patterns as selected features accordingly. There are two reasons that the research replaces $*$ by the empty string. First, Google does not provide any query involving regular expressions. Another reason is that in McCrae and Collier's experiment, it indicated that many of the patterns were inflexible and matched very rarely. Therefore, they simply allowed $*$ to match the empty string and discovered the modification greatly improved the synonymy classification. For these ten lexico-syntactic pattern-based features, (6) assesses semantic similarity between P and Q

$$\text{WebPattern}(P,Q) = \begin{cases} 0, & \text{if } H(Pattern) \leq c \\ \dfrac{H(Pattern)}{H(P \cap Q)}, & \text{otherwise.} \end{cases} \tag{6}$$

Here, the pattern indicates a particular lexico-syntactic pattern defined. Ten lexico-syntactic pattern-based features are derived and listed as follows: "of $P(Q)$," "$P(Q)$," "and $P(Q$," ",$P(Q$," "against $P(Q$," "prevalence of *P Q*," "patients with *P Q*," "*P* known as *Q*," "*P*/*Q*," and "*P*, *Q*." The page hit counts of the pattern *H(Pattern)* is normalized by the number of page counts via querying both P and Q applying the Google search engine. The rational behind is to eliminate biases for terms having different frequencies of appearances on web [23].

C. Feature Selection Strategy

Not all of the input features are equally weighted for classifiers. Some features can be redundant or irrelevant, in here, based on F-score [16] as the feature selection strategy, to rank the input features according to their importance. F-score is a simple technique which measures the discrimination of two sets of real numbers. Given training vectors $x_k, k = 1, 2, 3, \ldots, n$, if the number of positive and negative instances are $n+$ and $n-$,

$$\text{NGD}(P,Q) = \begin{cases} 0, & \text{if } H(P \cap Q) \leq c \\ \dfrac{\max(\log H(P), \log H(Q)) - \log H(P \cap Q)}{\log N - \min(\log H(P), \log H(Q))}, & \text{otherwise.} \end{cases} \tag{5}$$

the score of the ith feature is defined as

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (7)$$

where $\bar{x}_i, \bar{x}_i^{(+)}, \bar{x}_i^{(-)}$ are the average of the ith feature of all, positive, and negative datasets, respectively; $x_{k,i}^{(+)}$ is the ith feature of the kth positive instance, and $x_{k,i}^{(-)}$ is the ith feature of the kth negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the score is, the more discriminative this feature is. Therefore, 15 features are ranked according to the scores.

D. Machine Learning Algorithms

The section illustrates the leverage of a semantic similarity measure through integration of all the similarity scores as described in previous sections.

For each pair of concepts (P, Q), a feature vector F is created. First, by querying the Google search engine, page counts for P, Q, P AND Q, and ten lexico-syntactic patterns are collected. Second, 15 input features are calculated and ranked by F-scores. Finally, a 15-D feature vector F is yielded. Therefore, feature vectors for the synonymous pairs (positive training samples) as well as the nonsynonymous pairs (negative training samples) are generated. A two-class SVM with the feature vectors is trained. After applying synonymous and nonsynonymous pairs to train the SVM classifier, the semantic similarity between two given concepts can be calculated. Following the same approach, the training feature vectors can be generated, and a feature vector F0 for the given pair of concepts (P0, Q0) is created. The semantic similarity between them is measured. The semantic similarity between P0 and Q0 as the posterior probability, i.e., Prob(F0\synonymous), where the feature vector F0 belongs to the synonymous (positive) class.

Being a large-margin classifier, the output of an SVM is the distance from the decision hyperplane. However, this is not a calibrated posterior probability. The sigmoid functions are adapted to convert this distance into a posterior probability [17]. In the research, the libsvm 2.89 [16] toolbox including C-SVC and nu-SVC is utilized to perform the experiments.

V. Experiments

A. Datasets

Because of the lack of standard human rating benchmark datasets in biomedical domains, to evaluate the approaches, the dataset 1 [8], [24], [25] is applied containing 36 biomedical concept pairs as listed in the left three columns of Table IV. The human judgment scores (H) in this dataset are the mean values of the scores provided by reliable doctors. The dataset 2 [3] of 30 concept pairs from Pedersen *et al.*, annotated by nine medical index coders (Cod) and three physicians (Phy), are listed in the left four columns of Table V.

TABLE I
Features Ranked With F-scores

Rank	Feature	F(i)		Rank	Feature	F(i)
1	NGD	0.2751		9	WebJaccard	0.0347
2	WebPMI	0.237		10	of P (Q)	0.0185
3	, P (Q	0.1648		11	and P (Q	0.0093
4	P/Q	0.1632		12	against P (Q	0.0027
5	P(Q)	0.1606		13	patients with P Q	0.0017
6	P, Q	0.1585		14	P known as Q	0.0014
7	WebOverlap	0.1173		15	prevalence of P Q	0.0011
8	WebDice	0.0555				

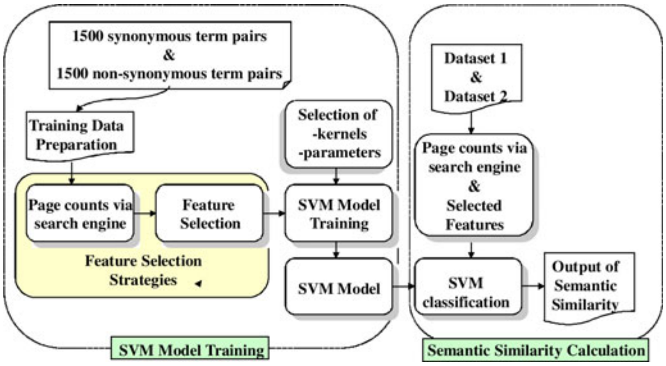
1500 synonymous term pairs
&
1500 non-synonymous term pairs
Training Data Preparation
Page counts via search engine
Feature Selection
Feature Selection Strategies
Selection of -kernels -parameters
SVM Model Training
SVM Model
SVM Model Training
Dataset 1
&
Dataset 2
Page counts via search engine
&
Selected Features
SVM classification
Output of Semantic Similarity
Semantic Similarity Calculation

Fig. 1. Detailed running procedures of methodologies.

B. Optimization and Calibration

1) Classifiers: Two classifiers, C-SVC and nu-SVC (support vector classifier: NU algorithm), are selected, based on four kernels [26]: linear kernel SVM, SVM-2 (polynomial kernel degree 2), SVM-3 (polynomial kernel degree 3), and radial basis function (RBF), respectively.

2) Number of Features: In the study, the ranked features based on the feature selection strategy, illustrated in Section IV-C, are presented in Table I in the descending order. In the table, the feature having the highest $F(i)$ value is NGD (0.2751), followed by a series of features such as WebPMI (0.237), "P(Q" (0.1648), "P/Q" (0.1632), "P(Q)" (0.1606), etc.

C. Running Procedures

A detailed running procedure of methodologies is presented in Fig. 1. In the SVM model training experiment, in order to determine the rankings of the features, initially, page hit counts, querying from Google, of 100 pairs of synonymous and nonsynonymous training samples are extracted and calculated according to the equations of the 15 defined features. Later, implementing the feature selection strategy as described previously, the obtained F-scores for each feature are averaged and ranked as indicated in Table I. In order to determine the optimal combination of features and training samples, the classifiers are trained, in the beginning, by input top two ranked features with the training samples, iterating from 100 pairs of synonymous and nonsynonymous until 1500 pairs with 100-pair increments in the ascending order. Afterward, the feature having the next higher rank was added, and the procedures were performed repeatedly until all 15 features are covered. Finally, the SVM models are generated.

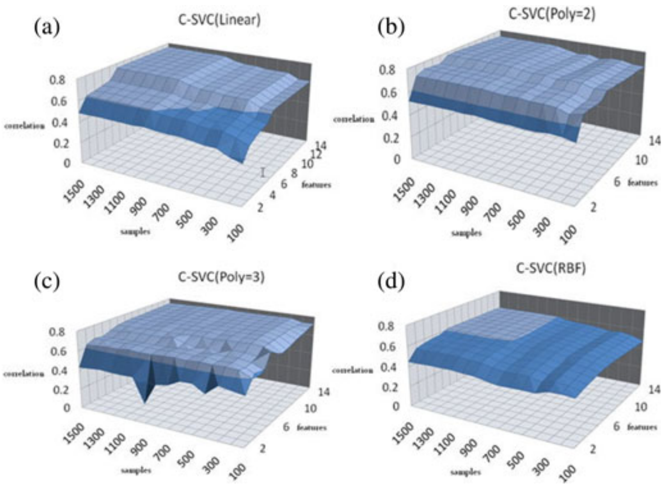

Fig. 2. Results for C-SVC models.

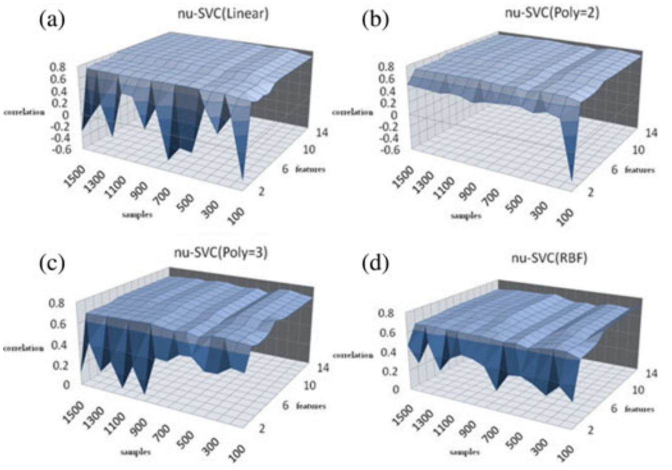

Fig. 3. Results for nu-SVC models.

After receiving the trained SVM model, input dataset 1 as testing data, and applying page counts via Google with selected features, the semantic similarity measures of the proposed method are produced and obtained. The optimal combinations of selected features and numbers of training samples are yielded as well. The outcomes are further implemented and validated against dataset 2.

VI. Results

The experiments are performed and executed via C-SVC, nu-SVC, based on four kernels (linear kernel SVM, SVM-2, SVM-3, as well as RBF) and the results, as depicted in Figs. 2, 3, respectively, are obtained.

In the diagrams, the X-coordinate represents the number of synonymous term pairs (positive training samples) and nonsynonymous term pairs (negative training samples), respectively. For example, the number “100” indicates 100 positive and 100 negative training samples; the numbers of term-pairs are ranged from 100 to 1500 with 100 increments. The Y-coordinate represents the numbers of features ranked in the descending order. The numbers of features are covered from 2 to 15. Initially, 100 positive and 100 negative training pairs as well as top two ranked features as input parameters apply into the C-SVC classifiers (with four kernels), and generate the trained models. After obtaining the models, dataset 1 is applied as testing data to compute the semantic similarity of each term pair in dataset 1. In addition, the correlations between the similarity measures and human judgment scores are calculated. The Z-coordinate indicates the individual correlations for the corresponding X and Y input parameters.

TABLE II
Correlations Versus Number of Samples and Features With Different Models

Model	Maximum correlation	Number of samples	Number of features
C-SVC(Linear)	0.758	1500	9
C-SVC(Poly=2)	0.776	1200	7
C-SVC(Poly=3)	0.759	300	13
C-SVC(RBF)	0.612	1100	10
nu-SVC(Linear)	0.798	900	7
nu-SVC(Poly=2)	0.766	300	11
nu-SVC(Poly=3)	0.736	300	12
nu-SVC(RBF)	0.743	100	11

The experimental results based upon C-SVC with linear kernel are summarized in Fig. 2(a). The maximum correlation coefficient of 0.758 is achieved with nine features and 1500 positive and negative training samples. Similarly, the results of polynomial degree 2 kernel are summarized in Fig. 2(b). In the diagram, the maximum correlation coefficient of 0.776 is achieved with seven features and 1200 samples each. The outcomes of polynomial degree 3 kernel are depicted in Fig. 2(c). The maximum correlation coefficient of 0.759 is achieved with 13 features and 300 samples each. Finally, the results of RBF kernel are displayed in Fig. 2(d). The maximum correlation coefficient of 0.612 is achieved with ten features and 1100 samples each.

The experimental results according to nu-SVC with four kernels are presented in Fig. 3. In the diagram, the definitions of the coordinates and parameters are identical as those in Fig. 2; the correlation coefficients are displayed.

The detailed experimental results of all models are listed, illustrated in Table II. In the table, it indicates that the linear kernel of nu-SVC has the highest correlation coefficient, 0.798, with 900 positive and negative samples each, having top seven ranked features. Similarly, in the C-SVC, SVM-2 has a correlation coefficient of 0.776. Apparently, in dataset 1, when higher degree kernels, i.e., SVM-2 and SVM-3 of nu-SVC are utilized, correlation with the human ratings decreases.

According to Table II, the optimal parameters of each model are obtained, e.g., in C-SVC (linear kernel) having 1500 positive and negative samples with top nine ranked features. Based on the parameters listed in Table II, each model, i.e., with optimal parameters applies the dataset 2 and calculates the results presented in Table III.

In the table, it indicates that the linear kernel of nu-SVC has the highest correlation coefficients, i.e., 0.705 (physician score) and 0.496 (coder score), respectively.

TABLE III
CORRELATIONS VERSUS DATASET 2 WITH PHYSICIAN SCORES AND CODER SCORES FOR INDIVIDUAL MODELS

Model	Dataset 2 (Physician)	Dataset 2 (Coder)	Number of samples	Number of features
C-SVC(Linear)	0.689	0.482	1500	9
C-SVC(Poly=2)	0.698	0.479	1200	7
C-SVC(Poly=3)	0.649	0.395	300	13
C-SVC(RBF)	0.388	0.171	1100	10
nu-SVC(Linear)	0.705	0.496	900	7
nu-SVC(Poly=2)	0.671	0.424	300	11
nu-SVC(Poly=3)	0.641	0.384	300	12
nu-SVC(RBF)	0.632	0.373	100	11

TABLE IV
DATASET 1 WITH HUMAN JUDGMENT SCORES AND PROPOSED SCORES

Concept 1	Concept 2	H	Proposed
Anemia	Appendicitis	0.031	0.697
Dementia	Atopic Dermatitis	0.062	0.371
Bacterial Pneumonia	Malaria	0.156	0.444
Osteoporosis	Patent Ductus Arteriosus	0.156	0.248
Amino Acid Sequence	Anti-Bacterial Agents	0.156	0.566
Acquired Immunodeficiency Syndrome	Congenital Heart Defects	0.062	0.210
Otitis Media	Infantile Colic	0.156	0.521
Meningitis	Tricuspid Atresia	0.031	0.256
Sinusitis	Mental Retardation	0.031	0.333
Hypertension	Kidney Failure	0.500	0.956
Hyperlipidemia	Hyperkalemia	0.156	0.568
Hypothyroidism	Hyperthyroidism	0.406	0.999
Sarcoidosis	Tuberculosis	0.406	0.996
Vaccines	Immunity	0.593	0.797
Asthma	Pneumonia	0.375	0.998
Diabetic Nephropathy	Diabetes Mellitus	0.500	0.950
Lactose Intolerance	Irritable Bowel Syndrome	0.468	0.883
Urinary Tract Infection	Pyelonephritis	0.656	0.991
Neonatal Jaundice	Sepsis	0.187	0.596
Sickle Cell Anemia	Iron Deficiency Anemia	0.437	0.686
Psychology	**Cognitive Science**	0.593	1.000
Adenovirus	Rotavirus	0.437	0.983
Migraine	Headache	0.718	1.000
Myocardial Ischemia	Myocardial Infarction	0.75	0.994
Hepatitis B	Hepatitis C	0.562	1.000
Carcinoma	Neoplasm	0.750	0.889
Pulmonary Valve Stenosis	Aortic Valve Stenosis	0.531	0.960
Failure To Thrive	Malnutrition	0.625	0.934
Breast Feeding	Lactation	0.843	0.976
Antibiotics	**Antibacterial Agents**	0.937	0.953
Seizures	Convulsions	0.843	1.000
Pain	Ache	0.875	0.830
Malnutrition	Nutritional Deficiency	0.875	0.923
Measles	Rubeola	0.906	1.000
Chicken Pox	Varicella	0.968	1.000
Down Syndrome	Trisomy 21	0.875	1.000
Correlation		1	0.798

TABLE V
DATASET 2 WITH PHYSICIAN & CODER SCORES AND PROPOSED SCORES

Concept 1	Concept 2	Phy	Cod	Proposed
Renal Failure	Kidney Failure	4	4	0.975
Heart	Myocardium	3.3	3	0.910
Stroke	Infarct	3	2.8	0.924
Abortion	Miscarriage	3	3.3	0.994
Delusion	Schizophrenia	3	2.2	0.500
Congestive Heart Failure	Pulmonary Edema	3	1.4	0.999
Metastasis	Adenocarcinoma	2.7	1.8	0.880
Calcification	Stenosis	2.7	2	0.748
Diarrhea	**Stomach Cramps**	2.3	1.3	1.000
Mitral Stenosis	Atrial Fibrillation	2.3	1.3	0.962
Chronic Obstructive Pulmonary Disease	**Lung Infiltrates**	2.3	1.8	0.349
Rheumatoid Arthritis	Lupus	2	1.1	0.998
Brain Tumor	Intracranial Hemorrhage	2	1.3	0.547
Carpal Tunnel Syndrome	Osteoarthritis	2	1.1	0.818
Diabetes Mellitus	Hypertension	2	1	1.000
Acne	Syringe	2	1	0.350
Antibiotic	Allergy	1.7	1.2	0.849
Cortisone	Total Knee Replacement	1.7	1	0.279
Pulmonary Embolus	**Myocardial Infarction**	1.7	1.2	0.940
Pulmonary Fibrosis	Lung Cancer	1.7	1.4	0.706
Cholangiocarcinoma	Colonoscopy	1.3	1	0.352
Lymphoid Hyperplasia	Laryngeal Cancer	1.3	1	0.241
Multiple Sclerosis	Psychosis	1	1	0.415
Appendicitis	Osteoporosis	1	1	0.570
Rectal Polyp	**Aorta**	1	1	0.296
Xerostomia	Alcoholic Cirrhosis	1	1	0.247
Peptic Ulcer Disease	Myopia	1	1	0.242
Depression	Cellulitis	1	1	0.376
Varicose Vein	**Entire Knee Meniscus**	1	1	NaN*
Hyperlipidemia	Metastasis	1	1	0.293
Correlation				0.705 0.496

*The input features have the threshold of page count C.

TABLE VI
RESULTS COMPARISON FOR CORRELATIONS USING SNOMED-CT ON DATASET 1*

SNOMED-CT	
Measure	Correl (Rank)
SemDist	0.735 (2)
Path length	0.586 (5)
Leacock & Chodorow	0.677 (4)
Wu & Palmer	0.686 (3)
Proposed	0.802 (1)

*34 Concept pairs out of 36 pairs.

VII. DISCUSSION

The concept pairs in datasets 1 and 2 are validated by applying the proposed semantic similarity measures. Results are presented in Tables IV and V, respectively. The proposed method earns the highest correlation of 0.798 in dataset 1, 0.705 in dataset 2 with physician's scores, and 0.496 in dataset 2 with coder's scores.

The last row in Table VI displays the correlation with human judgment scores of the proposed method (nu-SVC with seven features and 900 training samples) using dataset 1. Two terms shown in bold faced are excluded from SNOMED-CT terminology. Thus, the rest 34 pairs are applied to compute the correlations of measures in the table. The experimental

TABLE VII
CORRELATIONS USING SNOMED-CT ON DATASET 2

SNOMED-CT		
Measure	Physician Correlation (Rank)	Coder Correlation (Rank)
Path length	0.512 (4)	0.731 (2)
Leacock & Chodorow	0.358 (7)	0.497 (5)
Lin	0.522 (3)	0.565 (4)
Resnik	0.534 (2)	0.610 (3)
Jiang & Conrath	0.506 (5)	0.741 (1)
Vector (All sect, 1M notes)	0.436 (6)	0.497 (5)
Proposed	0.705 (1)	0.496 (6)

*28 Concept pairs out of 30 pairs.

TABLE VIII
CORRELATIONS USING MESH ON DATASET 1

MeSH			
Measure	Correl (Rank)	Measure	Correl (Rank)
SemDist	0.825 (1)	Resnik	0.718(7)
Path length	0.765 (5)	Li	0.705 (9)
Leacock & Chodorow	0.820 (2)	Lord	0.701 (10)
Wu & Palmer	0.811 (3)	Tversky	0.670 (11)
Lin	0.723 (6)	Rodriguez	0.690 (12)
Jiang & Conrath	0.710 (8)	Proposed	0.798 (4)

TABLE IX
CORRELATIONS USING MESH ON DATASET 2*

MeSH		
Measure	Physician Correlation (Rank)	Coder Correlation (Rank)
SemDist	0.666 (3)	0.863 (1)
Path length	0.627 (5)	0.744 (4)
Leacock & Chodorow	0.672 (2)	0.857 (2)
Wu & Palmer	0.652 (4)	0.794 (3)
Choi & Kim	0.560 (6)	0.724 (5)
Proposed	0.723 (1)	0.539 (6)

*25 Concept pairs out of the 30.

TABLE X
CORRELATIONS USING UMLS/SNOMED-CT ON DATASET 2

UMLS/SNOMED-CT		
Measure	Physician Correlation (Rank)	Coder Correlation (Rank)
Path length	0.231 (6)	0.320 (5)
Leacock & Chodorow	0.235 (5)	0.313 (6)
Lin	0.275 (4)	0.340 (4)
Resnik	0.420 (3)	0.532 (2)
Jiang & Conrath	0.485 (2)	0.603 (1)
Vector (All sect, 1M notes)	N/A	N/A
Proposed	0.681 (1)	0.481 (3)

*26 Concept pairs out of 30 pairs.

results comparing with four other measures: SemDist [8], path length [6], Leacock and Chodorow [27], and Wu and Palmer [28] are presented in Table VI.

Moreover, the results (nu-SVC with seven features and 900 training samples) for dataset 2 comparing with six other measures, path length [6], Leacock and Chodorow [27], Lin [29], Resnik [30], Jiang and Conrath [31], and Vector [3], are shown in Table VII. The proposed approach achieves the best correlation with physician scores comparing with those of the other six methods. However, the correlation with coder scores receives opposite outcomes. It can be concluded that the proposed approach is close to the physicians' ratings. The two term pairs chronic obstructive pulmonary disease/lung infiltrates and varicose vein/entire knee meniscus are excluded.

The preliminary results of UMLS-similarity [32] benchmark are performed on dataset 2 with 26 term pairs (four term pairs excluded) using the PAR/CHD relations in SNOMED-CT from the Unified Medical Language System (UMLS) applying to the corresponding measures in Table VII; the correlations are displayed in Table X. Apparently, having different term pairs to compute the correlations, the outcomes of the measures are different. The larger the numbers of term pairs involved, the higher the correlations turn out. However, for each measure, if a measure is closer to physician's findings in one table, it is also true in another table. Similarly, if a measure is closer to coder's score, it indicates the same results in both tables.

Table VIII indicates the results of correlations with human scores for the proposed approach (nu-SVC with seven features and 900 training samples) and other measures, applying dataset 1, experimented on MeSH. Comparing with 11 other measures, SemDist [8], path length [6], Leacock and Chodorow [27], Wu and Palmer [28], Lin [29], Jiang and Conrath [31], Resnik [30], Li *et al.* [33], Lord *et al.* [7], Tversky [34], and Rodriguez and Egenhofer [35], the correlation of the proposed approach achieves the fourth rank among the 12 methods.

Furthermore, Table IX indicates the results of correlations having physician and coder scores for the proposed approach (i.e., nu-SVC with seven features and 900 training samples) using dataset 2 comparing with other five methods. Because five term pairs shown in bold faced are excluded from MeSH, the 25 pairs are applied to compute the correlations of the measures in the table. The correlations of the proposed method for both physician and coder scores are, i.e., 0.723 and 0.539, respectively. In the table, it contains scores of the other five measures as well: SemDist [8], path length [6], Leacock and Chodorow [27], Wu and Palmer [28], and Choi and Kim [36]. The measures of the proposed method achieve the best correlation with physician scores and the sixth rank with coder scores.

Phrases such as known as, is a, part of, are examples of indications of various semantic relations. Some of such phrases are useful for capturing synonymous relationships. In the study, the feature, "P knows as Q", ranked 14, has been embedded in the ten lexico-syntactic patterns. However, the proposed method, i.e., the optimal SVM model, only involves the top seven features. Thus, identifying the exact set of words that convey the semantic relatedness between two concepts remains a challenging problem which requires further semantic analyses.

The test sets used in the research, i.e., datasets 1 and 2, are applied to establish relative performance of different measures and comparison among the measures for ontology based as well as corpus based approaches. In [37], it roughly divided the approaches of semantic similarity in the biomedical domain into knowledge based and distributional based methods. Knowledge based methods include path finding measures and intrinsic information content measures. The distributed measures utilize the distribution of concepts within a corpus that includes corpus information content and context vector methods. In [3], it classified [29], [31], and [30] as information content measures. On the other hand, path length [6], Leacock [27], and Wu [28] are classified as path finding measures. Pedersen *et al*. [3] also derived a context vector measure based on co-occurrence statistics of medical corpora that can be used as a measure of semantic relatedness. Furthermore, they found that the context vector measure correlates most closely with the physicians' ratings, while the path-based measures correlates most closely with the medical coders'. In the research in [14], the correlations, for information content based measures according to two corpora, the web, and a specific clinic corpus, are compared against physician judgments. Apparently, it can be assumed that the measures are close to the physician findings.

Hisham and Nguyen [8], [25] proposed a new ontology based technique for measuring semantic similarity in a single ontology as well as across multiple ontology in the biomedical domain. In the paper, the correlation between coders is stronger than that between physicians. They concluded and assumed that for the ontology-based method, the coders rating scores are more reliable than the physician rating scores.

The study has been adapted mainly on datasets provided by other works [3], [24]. To derive a reliable test set, the experiment needs to further interface with physicians or medical coders specializing in the same sub field of medicine to annotate domain-independent term pairs. Two criteria of the evaluation of the data are term pair integrity and user integrity [24] to exclude significant different results and getting good inter-annotator agreements. It is definitely a time consuming effort. Furthermore, the SVM models have been trained by input ranked features sequentially. The combinations of the features are not explored currently. The experiment can be enhanced in this direction as future work.

In addition, Bollegala *et al*. [9] explored semantic similarity measures by automatically extracted lexico-syntactic patterns from text snippets and four page-count based co-occurrence scores returned from Google search engine. In the paper, a maximum of 200 lexico-syntactic patterns were generated. A total of 204 dimensional feature patterns were trained on a two-class SVM applying 5000 synonymous and nonsynonymous word pairs. The paper also concluded that similarity measures based on lexico-syntactic patterns extracted from text snippets are more accurate than the four page-count based co-occurrence measures. The research reveals that correlation does not improve beyond 2500 positive and negative training examples. Therefore, they can conclude that 2500 examples are sufficient to leverage the proposed semantic similarity measure. The Web-Dice pattern has the highest kernel weight followed by a series of lexical pattern-based features.

Comparing with our proposed method, 15 patterns are defined with NGD [22] as the top ranked feature. The correlation does not improve beyond 1500 positive and negative training sample instances. We can conclude that our proposed method can determine term pairs' semantic similarity effectively and efficiently. It can be explored under real-time to achieve extracting medical terms across documents in HIS.

VIII. Conclusion

In the study, there are 15 input features defined for classifications, including five co-occurrence-based and ten lexico-syntactic pattern-based. In order to estimate the co-occurrence of the two concepts P and Q on the Web, the corresponding Google search engine's page counts are aggregated and calculated. The equation F-score is applied for ranking feature selections. Four kernels of C-SVC and nu-SVC are applied for classification models. The optimal parameters of each model are obtained by applying dataset 1 and validated by dataset 2. In addition, the medical datasets as experimental classifications include SNOMED-CT and MeSH standards.

The concept pairs in datasets 1 and 2 by applying the proposed semantic similarity measures are presented. In dataset 1, the proposed approach achieves the best correlation coefficient (0.802) for SNOMED-CT. In dataset 2, the approach obtains the best correlation coefficient (SNOMED-CT: 0.705; MeSH: 0.723) with physician scores comparing with measures of other methods. However, the correlation coefficients (SNOMED-CT: 0.496; MeSH: 0.539) with coder scores received opposite outcomes. In conclusion, the proposed method is close to physicians' ratings. In addition, the study provides the cornerstone to explore and implement a real-time application to extract semantic similarity terms across fully relevant information from NTUH electronic, free-text medical records, e.g., radiology reports and discharge notes.

References

[1] I. Neamatullah, M. M. Douglass, L. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, and G. D. Clifford, "Automated de-identification of free-text medical records," *BMC Med. Inf. Decision Making*, vol. 8, no. 32, pp. 1–17, Jul. 2008.

[2] T. Gong, C. L. Tan, T. Y. Leong, C. K. Lee, B. C. Pang, C. C. Tchoyoson Lim, Q. Tian, S. Tang, and Z. Zhang, "Text mining in radiology reports," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 815–820.

[3] T. Pedersen, S. Pakhomov, and S. Patwardhan, "Measures of semantic similarity and relatedness in the medical domain," *J. Biomed. Inf.*, vol. 40, no. 3, pp. 288–299, 2007.

[4] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder, "An environment for merging and testing large ontologies," in *Proc. 7th Int. Conf. Principles Knowl. Represent. Reason.*, Breckenridge, CO, USA, Apr. 12–15, 2000, pp. 483–493.

[5] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder, "The chimaera ontology environment," in *Proc. 17th Nat. Conf. Artif. Intell.*, Austin, TX, USA, Jul. 30, 2000, pp. 1123–1124.

[6] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 17–30, Jan.–Feb. 1989.

[7] P. Lord, R. Stevens, A. Brass, and C. Goble, "Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, 2003.

[8] H. Al-Mubaid and H. A. Nguyen, "Measuring semantic similarity between biomedical concepts within multiple ontologies," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 39, no. 4, pp. 389–398, Jul. 2009.

[9] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 757–766.

[10] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proc. 39th Annu. Meet. Assoc. Comput. Linguistics*, 2001, pp. 26–33.

[11] Y. Cao and H. Li, "Base noun phrase translation using web data and the EM algorithm," in *Proc. COLING*, 2002, pp. 127–133.

[12] T. Chklovski and P. Pantel, "VerbOcean: Mining the web for fine-grained semantic verb relations," in *Proc. Conf. Empirical Methods Natural Language Process.*, Barcelona, Spain, 2004, pp. 33–40.

[13] P. Nakov and M. Hearst, "A study of using search engine page hits as a proxy for n-gram frequencies," presented at the Recent Adv. Natural Language Process., Borovets, Bulgaria, 2005.

[14] D. Sánchez, M. Batet, and A. Valls, "Computing knowledge-based semantic similarity from the web: An application to the biomedical domain," in *Proc. 3rd Int. Conf. Knowl. Sci. Eng. Manage.*, 2009, pp. 17–28.

[15] G. Pirró and J. Euzenat, "A feature and information theoretic framework for semantic similarity and relatedness," in *Proc. 9th Int. Semantic Web Conf. Semantic Web*, 2010, pp. 615–630.

[16] Y. W. Chen and C. J. Lin, "Combining SVMs with various feature selection strategies," *Stud. Fuzziness Soft Comput.*, vol. 207, pp. 315–324, 2006.

[17] J. C. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers,.* Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[18] SNOMED-CT site [Online]. Available: http://www.ihtsdo.org/snomed-ct/snomed-ct0

[19] MeSH site [Online]. Available: http://www.nlm.nih.gov/mesh/MBrowser.html

[20] Medicine Net.com, Health and Medical Information Produced by Doctors [Online]. Available: http://www.medterms.com/script/main/hp.asp

[21] Synonyms.net [Online]. Available: http://www.synonyms.net

[22] R. L. Cilibrasi and P. M. B. Vitanyi, "The google similarity distance," presented at the IEEE ITSOC Inf. Theory Workshop, Rotorua, New Zealand, Aug. 29–Sep. 1, 2005.

[23] J. McCrae and N. Collier, "Synonym set extraction from the biomedical literature by lexical pattern discovery," *BMC Bioinf.*, vol. 9, 2008.

[24] A. Hliaoutakis, "Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline," M.S. thesis, Tech. Univ. Crete, Crete, Greece, 2005.

[25] A.-M. Hisham and H. A. Nguyen, "A cluster-based approach for semantic similarity in the biomedical domain," in *Proc. 28th IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, New York, USA, Aug. 30–Sep. 3, 2006, pp. 2713–2717.

[26] LIBSVM—A library for support vector machines [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[27] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed. Cambridge, MA, USA: MIT Press, 1998, pp. 305–332.

[28] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proc. 32nd Annu. Meet. Assoc. Comput. Linguistics*, 1994, pp. 133–138.

[29] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Mach. Learn.*, Madison, WI, USA, 1998, pp. 296–304.

[30] P. Resnik, "WordNet and class-based probabilities," C. Fellbaum, Ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998, pp. 239–263.

[31] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. 10th Int. Conf. Res. Comput. Linguistics*, Taipei, Taiwan, 1997, pp. 19–33.

[32] B. T. McInnes, T. Pedersen, and S. V. S. Pakhomov, "UMLS-interface and UMLS-similarity: Open source software for measuring paths and semantic similarity," in *Proc. Amer. Med. Inf. Assoc. Symp.*, San Francisco, CA, USA, 2009, pp. 431–435.

[33] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 871–882, Jul.–Aug. 2003.

[34] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977.

[35] A. Rodriguez and M. J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 2, pp. 442–456, Mar.–Apr. 2003.

[36] I. Choi and M. Kim, "Topic distillation using hierarchy concept tree," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, 2003, pp. 371–372.

[37] V. N. Garla and C. Brandt, "Semantic similarity in the biomedical domain: An evaluation across knowledge sources," *BMC Bioinf.*, vol. 13, no. 1, p. 261, Oct. 2012.

Sheau-Ling Hsieh received three M.S. degrees as well as a Ph.D. degree from the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA, in 1998. She is currently an Associate Professor with National Chiao Tung University (NCTU), Hsinchu, Taiwan, offering courses in EE/CS International Graduate Program.

She worked for Intel, Unisys, Bull, Mitre, Philips, and start-ups in USA, was in charge of R&D before going to Japan, and later joined NCTU. Her research has been focused on networking related device drivers, telecommunication protocols, as well as service oriented architecture, middleware frameworks for e-commerce applications, especially applying to hospital information systems and renewable natural resources & ecological environment. She is also a member of Disaster Prevention & Water Environment Research Center, NCTU.

Wen-Yung Chang received the M.S. degree in 2009. He was with the College of Computer Science, National Chiao Tung University.

Chi-Huang Chen received the Ph.D. degree in 2010. He was with the Department of Electrical Engineering, National Taiwan University.

Yung-Ching Weng received the Ph.D. degree from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, in 2010. He is with the Information Systems Office, National Taiwan University Hospital, Taipei, as a Senior Engineer.