

國立交通大學

資訊科學與工程研究所

碩士論文

生醫領域語意相似度測量

Semantic Similarity Measurement in Biomedical Domain

研究生：張文勇

指導教授：謝筱齡/林正中 教授

中華民國九十八年七月

生醫領域語意相似度測量  
Semantic Similarity Measurement in Biomedical Domain

研究生：張文勇

Student : Wen-Yung Chang

指導教授：謝筱齡/林正中

Advisor : Sheau-Ling Hsieh/ Cheng-Chung Lin

國立交通大學  
資訊科學與工程研究所  
碩士論文

A Thesis  
Submitted to Institute of Computer Science and Engineering  
College of Computer Science  
National Chiao Tung University  
in partial Fulfillment of the Requirements  
for the Degree of  
Master  
in  
Computer Science

July 2008

Hsinchu, Taiwan, Republic of China

中華民國九十八年七月

學生：張文勇

指導教授：謝筱齡  
林正中

國立交通大學資訊科學與工程所 碩士班

## 摘要

語意相似性度計算在信息檢索和自然語言處理領域扮演重要的角色。在本文中，我們提出了一種基於網頁數的語意相似性度計算方法並應用到生物醫學領域。以往的研究中語意網相關的應用已經使用了各種語意相似性度計算的方法。儘管語意相似性度計算應用範圍甚廣，但是測量兩個詞之間的語意相似性度仍然是一個具有挑戰性的任務。本文提出的方法利用搜尋引擎傳回的網頁數來計算語意相似性度。給予兩個詞  $P$  和  $Q$ ，利用網頁數的查詢  $P$  和  $Q$  和  $P \text{ AND } Q$  以及所定義的公式作為計算整合我們提出的一種新方法使用一些句法查詢其出現的網頁數來計算語意相似性度。將這些不同的相似分數分別使用支持向量機和決策樹學習，再計算其出現於同義字類別的機率作為語意的相似性度。兩組數據實驗結果顯示，在第一組(A. Hliaoutakis 所提出)可以達到 0.798 的相關係數，在第二組(T. Pedersen 等人所提出)以醫生的分數為基準可以達到 0.705 的相關係數，以醫學專業人員的分數為基準可以達到 0.496 的相關係數。

關鍵字：生醫術語、語意相似性度、網路探勘

# Semantic Similarity Measurement in Biomedical Domain

student : Wen-Yung Chang

Advisors : Dr. Sheau-Ling Hsieh  
Dr. Cheng-Chung Lin

## Abstract

Semantic similarity measure plays an important role in Information Retrieval and Natural Language Processing. In this paper we propose a page-count-based semantic similarity measure and apply it into the biomedical domain. Previous work in semantic web related applications have used various semantic similarity measures. Despite the usefulness of these applications, measuring semantic similarity between two terms remains a challenging task. The proposed method exploits page counts returned by the Web search engine. We define various similarity scores for two given terms  $P$  and  $Q$ , using the page counts for the queries  $P$ ,  $Q$  and  $P AND Q$ . Moreover, we propose a novel approach to compute semantic similarity based upon lexico-syntactic patterns using page counts. The different similarity scores are integrated with support vector machines and decision tree classifier models, to leverage a robustness of the measures. Experimental results achieve a correlation coefficient of 0.798 on the dataset provided by A. Hliaoutakis, 0.705 on the dataset provide by T. Pedersen et al with physician scores and 0.496 with expert scores, respectively.

Keywords: biomedical terminology, semantic similarity, web minning

## 致謝

首先，我想要感謝我的父母與家人在這一路上的支持與鼓勵，我才有辦法順利完成我的碩士學業，其次，要感謝我的兩位指導教授謝筱齡與林正中老師，在兩年的碩士生活中，謝筱齡老師在各方面的指導與幫忙，讓我受益不少，林正中老師提供了良好的實驗室環境與設備，讓我能完成我的學業。同時我也要感謝台大計算機中心陳啟煌學長，在討論中給我的寶貴意見。另外我也要謝謝資訊科學與工程研究所的林進燈教授，儘管您在電機與資訊領域扮演著教父一樣的角色，但您總是非常有耐心並且願意回答我所有基礎的問題。也因為您在學術界的深度與廣度讓我了解到做學問是刺激、富挑戰性又可以從中學習到很多東西的一個過程。

最後，感謝上天，總是在遙遠的天邊靜靜的看著我，在我意志消沉的時候派遣我生命中的貴人來幫助我；在我得意忘形的時候讓我摔跤，這樣我才會繼續虛心的學習。我了解我付出的是那麼少，得到的卻是那樣多，我應該要珍惜我擁有的一切。學海無涯，讀了越多東西才越清楚自己的學識有多麼的狹隘，每當我因為了解新東西而沾沾自喜時，卻又發現真正的學者總是謙沖而自牧的。



# Contents

摘要 .....	i
Abstract .....	ii
致謝 .....	iii
Contents .....	iv
List of Tables .....	v
List of Figures .....	v
<b>Chapter 1 Introduction .....</b>	<b>1</b>
<b>Chapter 2 Background and Related Work.....</b>	<b>4</b>
2.1 Google AJAX Search API .....	4
2.2 Support Vector Machine .....	4
2.3 Decision Tree .....	13
2.4 SNOMED-CT .....	14
2.5 MeSH.....	15
2.6 Semantic Similarity Measurement Methods .....	15
2.6.1 Edge-Counting Measures .....	15
2.6.2 Information Content Measures .....	16
2.6.3 Feature-Based Measures.....	17
2.6.4 Hybrid Measures .....	18
<b>Chapter 3 Methodology .....</b>	<b>19</b>
3.1 Sample Construction .....	19
3.2 Feature Definitions .....	21
3.3 Feature Selection Strategy .....	25
3.4 Support Vector Machine Model.....	27
3.5 Decision Tree Model .....	28
<b>Chapter 4 Experiment Results .....</b>	<b>31</b>
4.1 Datasets.....	31
4.2 Experiment Environment.....	34
4.3 Parameter Optimization.....	34
4.3.1 Classifier Models .....	34
4.3.2 Number of Features and Training Samples .....	34
4.4 Results .....	42
4.5 Comparison.....	46
<b>Chapter 5 Conclusions and Future Work.....</b>	<b>53</b>
5.1 Conclusions .....	53
5.2 Future Work.....	53
<b>References.....</b>	<b>54</b>
<b>Appendix .....</b>	<b>57</b>

## List of Tables

Table 3.1: Lexico-syntactic patterns.....	24
Table 4.1: Dataset 1 of 36 biomedical concept pairs.....	32
Table 4.2: Dataset 2 of 30 biomedical concept pairs sorted in the order of the averaged physician’s scores .....	33
Table 4.3: Features with highest F-scores .....	35
Table 4.4: Correlation vs. No of samples and features with different models.....	41
Table 4.5: Correlation vs. Dataset 1 and Dataset 2 with physician scores and expert scores of different models .....	45
Table 4.6: Dataset 1 with human similarity scores and proposed scores .....	46
Table 4.7: Dataset 2 with human similarity scores and proposed scores. ....	47
Table 4.8: Absolute correlations with human scores using SNOMED-CT on dataset 1. ....	49
Table 4.9: Absolute correlations with human scores using SNOMED-CT on dataset 2.....	49
Table 4.10: Absolute correlations with human scores using MeSH on dataset 1.....	50
Table 4.11: Absolute correlations with human scores using MeSH on dataset 2. ....	52

## List of Figures

Figure 2.1: Maximum-margin hyperplane in linear separable case .....	6
Figure 2.2: Maximum-margin hyperplane in nonlinear separable case .....	8
Figure 2.3: Mapping the training data into a high dimensional feature space by $\phi(\mathbf{x})$ where a linear separation is made, corresponding to a nonlinear separation in the original input space	16
Figure 2.4: Decision tree .....	14
Figure 3.1: MedicineNet.com website.....	20
Figure 3.2: synonyms.net website .....	21
Figure 3.3: Both features of this data have low scores as in equation 3.6 the denominator is much larger than the numerator.....	27
Figure 3.4: Support vector machine model flow chart .....	27
Figure 3.5: Decision tree model flow chart .....	29
Figure 4.1: Correlation vs. No of features and training samples using C-SVC with linear kernel	36
Figure 4.2: Correlation vs. No of features and training samples using C-SVC with polynomial degree=2 kernel .....	36
Figure 4.3: Correlation vs. No of features and training samples using C-SVC with polynomial degree=3 kernel .....	37
Figure 4.4: Correlation vs. No of features and training samples using C-SVC with RBF kernel	36

Figure 4.5: Correlation vs. No of features and training samples using nu-SVC with linear kernel 38

Figure 4.6: Correlation vs. No of features and training samples using nu-SVC with polynomial degree=2 kernel .....39

Figure 4.7: Correlation vs. No of features and training samples using nu-SVC with polynomial degree=3 kernel .....39

Figure 4.8: Correlation vs. No of features and training samples using nu-SVC with RBF kernel 40

Figure 4.9: Correlation vs. No of features and training samples using decision tree .....41

Figure 4.10: Correlation vs. Classifiers of dataset 1 with human scores.....43

Figure 4.11: Correlation vs. Classifiers of dataset 2 with physician scores .....44

Figure 4.12: Correlations vs. Classifiers of dataset 2 with expert scores .....45





# Chapter 1 Introduction

With the rapid growth of today's internet, in order to facilitate the management and search, various information documents has transformed into electronic documents. All types of information documents on the internet increased the difficulty of information retrieval. Research of semantic similarity between concepts has been an integral part of information retrieval and natural language processing.

The existence of semantic equivalence classes between lexical items in English makes it highly desirable to use thesauri of synonymous concepts for document retrieval (DR) and information retrieval (IR) applications. The issue is particularly acute in the biomedical domain due to stringent completeness requirements on such as patient cohort identification. We believe that measures of semantic similarity can improve the performance of such systems. For example, a user's query for "congestive heart failure" could be expanded to include the semantically similar terms of cardiac decompensation, pulmonary edema, ischemic cardiomyopathy and volume overload. Clearly, pulmonary edema does not denote the same or even a similar disorder as congestive heart failure but under the patient cohort identification conditions it could be considered as an equivalent search term.

Semantic similarity refers to human judgments of the degree to which a given pair of concepts. Measures of semantic similarity are automatic techniques that attempt to imitate human judgments of relatedness. Semantic similarity measures are classified into two main categories such as ontology-based and corpus-based. The first class is to measure the semantic similarity between two concepts  $c_1$ ,  $c_2$  by calculate the distance between the concept nodes in the ontology tree or hierarchy [1, 2]. The second class of techniques measures the difference of information content of the two concepts as a function of their probability of occurrence in a corpus. In this class, the techniques use machine learning, rule-based, statistical-based or other corpus-based approaches [2, 3, 4]. The corpus-based approach uses the information

available in the corpus to measure similarity between concepts or entities. In our research we use the corpus-based technique to measure the semantic similarity between concepts.

By using corpus-based approach how many corpus is an important issue in many Natural Language Processing (NLP) tasks. In 2001, (Banko & Brill 01) advocated for the creative using very large corpus as an alternative to sophisticated algorithms. They demonstrated the idea on a lexical disambiguation problem. The problem was to choose which of 2-3 commonly confused concepts were appropriate for a given context. They show that even using a very simple algorithm, the results continue improving log-linearly with more training data, even out to a billion concepts. They conclude that getting more data may be a better idea than fine tuning algorithms. The Web is providing unprecedented access to the information as well as interacting with people's daily lives. Today, the obvious source of largest data is the web.

Using the web as training and testing corpus is attracting ever-increasing attention. The web has been used as a corpus for a variety of NLP tasks such as machine translation (Grefenstette 98; Resnik 99; Cao & Li 02; Way & Gough 03), question answering: (Dumais et al. 02; Soricut & Brill 04), word sense disambiguation (Mihalcea & Moldovan 99; Rigau et al. 02; Santamaría et al. 03; Zahariev 04), extraction of semantic relations, (Chklovski & Pantel 04; Idan Szpektor & Coppola 04; Shinzato & Torisawa 04), anaphora resolution: (Modjeska et al. 03), prepositional phrase attachment: (Volk 01; Calvo & Gelbukh 03), language modeling: (Zhu & Rosenfeld 01; Keller & Lapata 03), semantic similarity (Danushka Bollegala & Yutaka Matsuo 07). In our research we proposed a method for semantic similarity measurement between concepts using web search engine and apply it into biomedical domain.

The rest of this thesis is organized as follows:

In chapter 2 we provide the necessary technical background and analysis of related work.

In chapter 3 we are going to introduce our similarity measurement methodology.

In chapter 4 we make the experiments and present the experiment results.

In chapter 5 we highlight the conclusions from the study, and propose future work.



## Chapter 2 Background and Related Work

In this chapter, the first three sections are going to report some technical background related to our research, we will introduce Google AJAX Search API in chapter 2.1, supervised learning methods used for classification and regression of support vector machine (SVM) in chapter 2.2 and decision tree classification method in chapter 2.3, for the following two sections described the SNOMED-CT ontology and the MeSH ontology in chapter 2.4 and chapter 2.5 respectively. Finally we introduce four major categories of ontology-based semantic similarity measurement methods in chapter 2.5.

### 2.1 Google AJAX Search API

The Google AJAX Search API is made up of four major classes of components:

The first class `google.search.SearchControl` provides the user interface and coordination over numbers of searcher objects, each searcher object is designed in order to perform searches and return a specific class of results.

The second class `google.search.Search` is also the base class which all "searchers" inherit. It defines the interface that all searcher services have to implement.

The third class `GResult` is also a base class that encapsulates the search results produced by the searcher objects.

The last class is `google.search.SearcherOptions`, this class configures the behavior of searcher objects when we add to a search control.

The detail discussion of how to use the Google AJAX Search API command is in appendix A.

### 2.2 Support Vector Machine

Support vector machine (SVM) is a supervised learning method used for classification and regression. SVM has been using widely because of its high generalization ability and the wide area of applications.

In two-class classification problem, a training set  $S = \{(\mathbf{x}_k, y_k)\}, k = 1, 2, \dots, n$ .  $\mathbf{x}_k$  describes the input patterns in  $d$ -dimensional feature space,  $\mathbf{x}_k \in R^d$ . The class labels  $y_k$  confirms as responses of  $\mathbf{x}_k$  from either of the two class, and are assigned with a value of +1 or -1. Our purpose is to find the hyperplane of the following equation

$$y_H(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b, \quad (2.3)$$

define the pair  $(\mathbf{w}, b)$ , such as the linear classifier

$$y(\mathbf{x}) = \text{sign}[\mathbf{w}^T \mathbf{x} + b], \quad (2.4)$$

where  $\mathbf{w} \in R^d$  and  $b \in R$ . When the data of the two classes are separable then it satisfies

$$\mathbf{w}^T \mathbf{x}_k + b \geq 1, \text{ if } y_k = +1, \quad (2.5)$$

$$\mathbf{w}^T \mathbf{x}_k + b \leq -1, \text{ if } y_k = -1. \quad (2.6)$$

If the set  $S$  is linear separable, the hyperplane can be combined into one inequality as follows

$$y_k[\mathbf{w}^T \mathbf{x}_k + b] \geq 1, \text{ for } k = 1, 2, \dots, n. \quad (2.7)$$

For the linear separable set  $S$ , we would like to find the hyperplane with largest margin. In other words, we would like the distance between two classes of training data as large as possible. The distance  $d(\mathbf{w}, b | \mathbf{x})$  from a point  $\mathbf{x}$  to the hyperplane  $(\mathbf{w}, b)$  is

$$d(\mathbf{w}, b | \mathbf{x}) \triangleq \frac{|\mathbf{w} \cdot \mathbf{x} + b|}{\|\mathbf{w}\|}. \quad (2.8)$$

The margin  $M$  is given by

$$\begin{aligned} M(\mathbf{w}, b) &= \min_{\{\mathbf{x}_k: y_k = -1\}} d(\mathbf{w}, b | \mathbf{x}_k) + \min_{\{\mathbf{x}_k: y_k = +1\}} d(\mathbf{w}, b | \mathbf{x}_k) \\ &= \min_{\{\mathbf{x}_k: y_k = -1\}} \frac{|\mathbf{w} \cdot \mathbf{x}_k + b|}{\|\mathbf{w}\|} + \min_{\{\mathbf{x}_k: y_k = +1\}} \frac{|\mathbf{w} \cdot \mathbf{x}_k + b|}{\|\mathbf{w}\|} \end{aligned} \quad (2.9)$$

$$\begin{aligned}
&= \frac{1}{\|\mathbf{w}\|} \left( \min_{\{x_k: y_k = -1\}} |\mathbf{w} \cdot \mathbf{x}_k + b| + \min_{\{x_k: y_k = +1\}} |\mathbf{w} \cdot \mathbf{x}_k + b| \right) \\
&= \frac{2}{\|\mathbf{w}\|}.
\end{aligned}$$

In Figure 2.1, the optimal hyperplane is given by maximizing the margin  $M$ , subject to the constraints of equation 2.4.

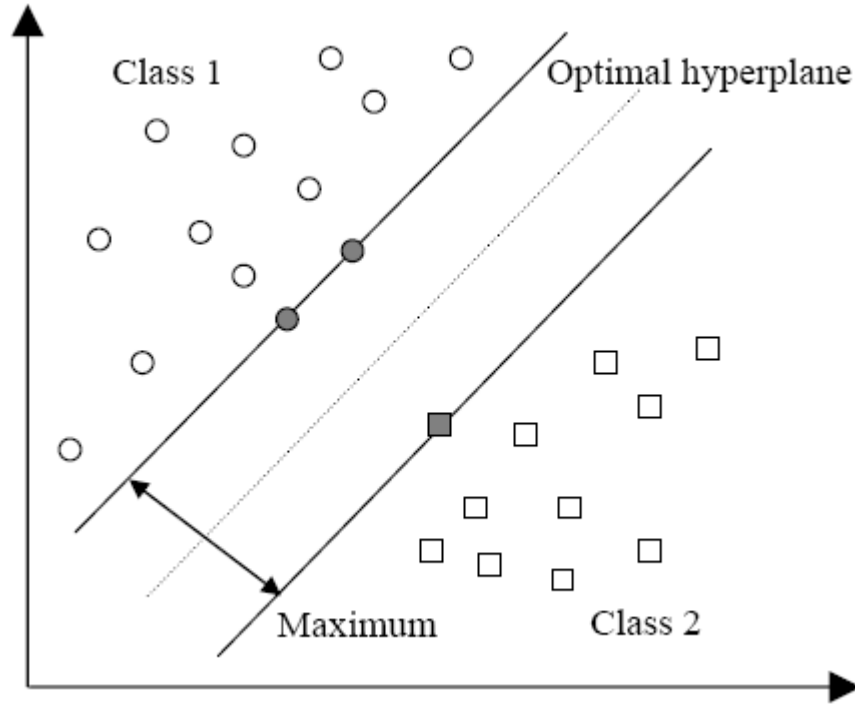


Figure 2.1: Maximum-margin hyperplane in linear separable case

The optimal hyperplane can be found by solving the following equation

$$\text{minimize } J_p(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w}, \tag{2.10}$$

$$\text{subject to } y_k [\mathbf{w}^T \mathbf{x}_k + b] \geq 1, \text{ for } k = 1, 2, \dots, n.$$

Searching the optimal hyperplane is a quadratic programming (QP) problem. This problem can be solved by constructing a Lagrangian

$$L(\mathbf{w}, b; \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{k=1}^n \alpha_k (y_k [\mathbf{w}^T \mathbf{x}_k + b] - 1), \tag{2.11}$$

where  $\alpha_k \geq 0$  are Lagrange multipliers. In order to find the saddle point we need to minimize this function over  $\mathbf{w}$  and  $b$  and maximize it over the nonnegative Lagrange multipliers  $\alpha_k \geq 0$ . At the saddle point, obtains

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{k=1}^n \alpha_k y_k \mathbf{x}_k = 0, \quad (2.12)$$

$$\frac{\partial L}{\partial b} = \sum_{k=1}^n \alpha_k y_k = 0. \quad (2.13)$$

Substitute equations 2.12 and 2.13 into 2.11, becomes the following QP problem as the dual problem

$$\text{maximize } J_D(\alpha) = -\frac{1}{2} \sum_{k,l=1}^n y_k y_l \mathbf{x}_k^T \mathbf{x}_l \alpha_k \alpha_l + \sum_{k=1}^n \alpha_k, \quad (2.14)$$

$$\text{subject to } \sum_{k=1}^n \alpha_k y_k = 0.$$

The Karush-Kuhn-Tucker (KKT) theorem plays an important role of SVM. Thus solving the SVM problem is equivalent to find the solution under KKT condition. According to this, the solution of equation 2.14 has the equality

$$\alpha_k (y_k (\mathbf{w} \cdot \mathbf{x}_k + b) - 1) = 0, \text{ for } k = 1, 2, \dots, n. \quad (2.15)$$

To construct the optimal hyperplane  $\mathbf{w} \cdot \mathbf{x} + b$ , from equation 2.12 it follows that

$$\mathbf{w} = \sum_{k=1}^n \alpha_k y_k \mathbf{x}_k, \quad (2.16)$$

and the scalar  $b$  can be determined from the KKT conditions of equation 2.15, such that the linear SVM classifier takes the form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{k=1}^n \alpha_k y_k \mathbf{x}_k^T \mathbf{x} + b\right). \quad (2.17)$$

At the same time, each training sample  $\mathbf{x}_k$  is associated with Lagrange coefficient  $\alpha_k$ . The sample whose coefficient  $\alpha_k$  is nonzero is called support vector.

In the previous section, the SVM solution is to a linear separable classification problem. However, most of cases are not linear separable where is an example in Figure 2.2.

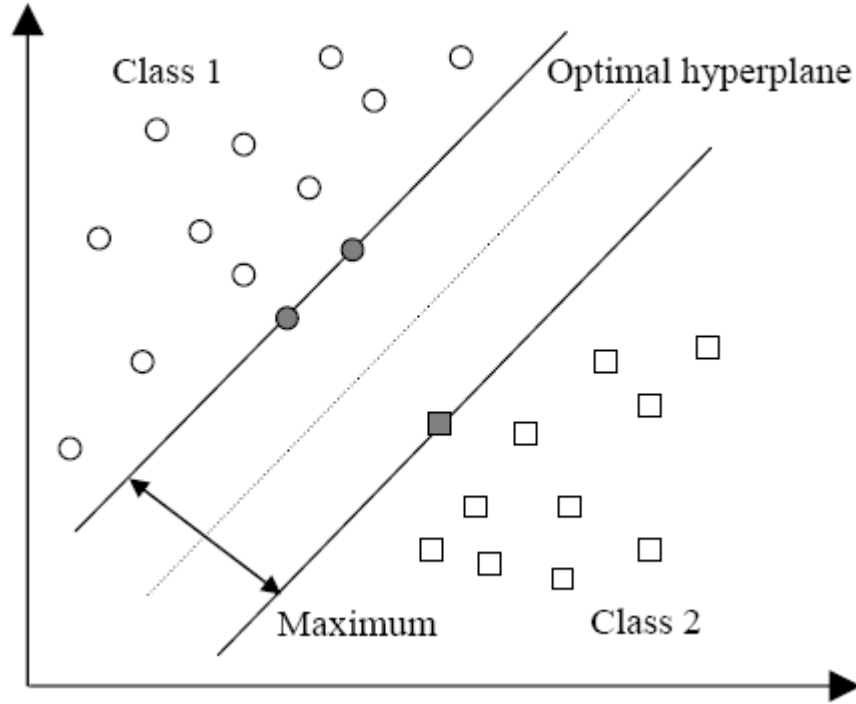


Figure 2.2: Maximum-margin hyperplane in nonlinear separable case

To solve the problem which is not linear separable in  $R^d$ , it is done by taking additional slack variables in the problem formulation. In order to tolerate misclassification, we have to modify the set of inequality equation 2.7 into

$$y_k[\mathbf{w}^T \mathbf{x}_k + b] \geq 1 - \xi_k, \text{ for } k = 1, 2, \dots, n, \quad (2.18)$$

where slack variable  $\xi_k \geq 0$ . In the primal weight space the optimization problem becomes

$$\text{minimize } J_p(\mathbf{w}, \xi) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + c \sum_{k=1}^n \xi_k, \quad (2.19)$$

$$\text{subject to } \begin{cases} y_k[\mathbf{w}^T \mathbf{x}_k + b] \geq 1 - \xi_k, & \text{for } k = 1, 2, \dots, n, \\ \xi_k \geq 0 \end{cases}$$

where  $c$  is a real constant. On the analogy of what was done for the separable case, the



solution to equation 2.19 is reduced to a QP optimization problem

$$\text{maximize } J_D(\alpha) = -\frac{1}{2} \sum_{k,l=1}^n y_k y_l \mathbf{x}_k^T \mathbf{x}_l \alpha_k \alpha_l + \sum_{k=1}^n \alpha_k, \quad (2.20)$$

$$\text{subject to } \begin{cases} \sum_{k=1}^n \alpha_k y_k = 0 \\ 0 \leq \alpha_k \leq c, k = 1, 2, \dots, n, \end{cases}$$

and the KKT conditions are defined as

$$\alpha_k (y_k (\mathbf{w} \cdot \mathbf{x}_k + b) - 1 + \xi_k) = 0, \text{ for } k = 1, 2, \dots, n, \quad (2.21)$$

$$(c - \alpha_k) \xi_k = 0, \text{ for } k = 1, 2, \dots, n. \quad (2.22)$$

The training data corresponding to non-zero  $\alpha_k$  value is called support vector, but there are two types of support vector in non-separable case. In the case  $0 < \alpha_k < c$ , the corresponding support vector  $\mathbf{x}_k$  satisfies the equalities  $y_k (\mathbf{w} \cdot \mathbf{x}_k + b) = 1$  and  $\xi_k = 0$ . This has no difference with separable case. In another case  $\alpha_k = c$ , the corresponding  $\xi_k$  is not null and the corresponding support vector  $\mathbf{x}_k$  does not satisfy equation 2.18. We refer such support vector as error. The point  $\mathbf{x}_k$  corresponding with  $\alpha_k = 0$  is classified correctly and far away from the decision margin.

To extend linear SVM classifiers to nonlinear SVM classifiers is straightforward. The case where a linear boundary is inappropriate to the SVM can map the input vector  $\mathbf{x}$  into a high dimensional feature space  $Z$ . In Figure 2.3, a construction of the linear separating hyperplane is done in this high dimensional feature space, after a nonlinear mapping  $\varphi(\mathbf{x})$  of the input data to the feature space.

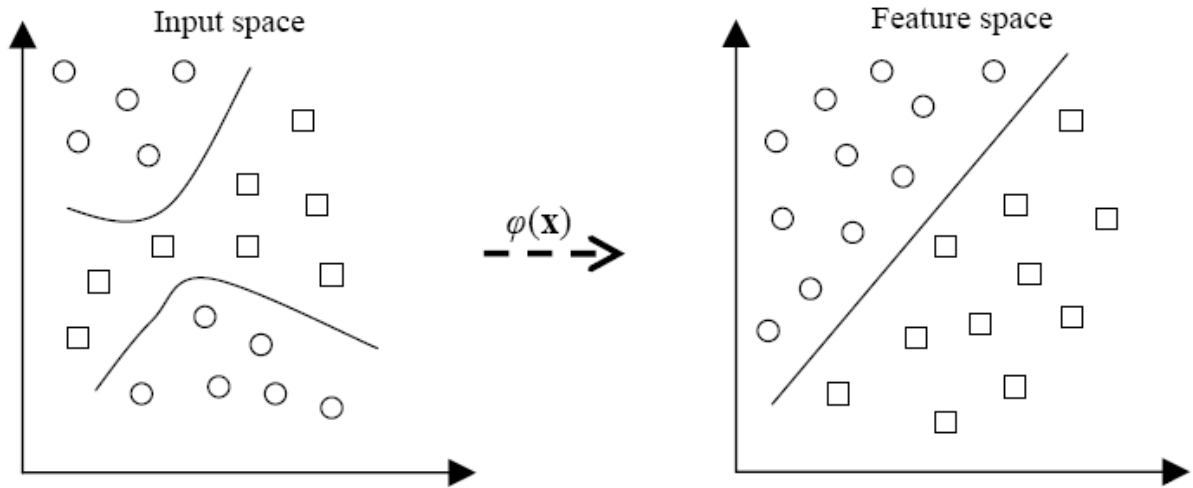


Figure 2.3: Mapping the training data into a high dimensional feature space by  $\varphi(\mathbf{x})$  where a linear separation is made, corresponding to a nonlinear separation in the original input space

The optimization problem of equation 2.19 becomes

$$\text{minimize } J_p(\mathbf{w}, \xi) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + c \sum_{k=1}^n \xi_k, \quad (2.23)$$

$$\text{subject to } \begin{cases} y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) + b] \geq 1 - \xi_k \\ \xi_k \geq 0 \end{cases}, \text{ for } k = 1, 2, \dots, n.$$

Construct the Lagrangian

$$L(\mathbf{w}, b, \xi, \alpha, v) \quad (2.24)$$

$$\begin{aligned} &= J(\mathbf{w}, \xi) - \sum_{k=1}^n \alpha_k (y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) + b] - 1 + \xi_k) - \sum_{k=1}^n v_k \xi_k \\ &= \frac{1}{2} \mathbf{w}^T \mathbf{w} + c \sum_{k=1}^n \xi_k - \sum_{k=1}^n \alpha_k y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) + b] + \sum_{k=1}^n \alpha_k - \sum_{k=1}^n \alpha_k \xi_k - \sum_{k=1}^n v_k \xi_k, \end{aligned}$$

with Lagrange multipliers  $\alpha_k \geq 0, v_k \geq 0$  for  $k = 1, 2, \dots, n$ . The solution is given by the

saddle point of the Lagrangian

$$\max_{\alpha, v} \min_{\mathbf{w}, b, \xi} L(\mathbf{w}, b, \xi, \alpha, v). \quad (2.25)$$

which obtains

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = 0 \rightarrow \mathbf{w} = \sum_{k=1}^n \alpha_k y_k \varphi(\mathbf{x}_k) \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^n \alpha_k y_k = 0 \\ \frac{\partial L}{\partial \xi_k} = 0 \rightarrow c - \alpha_k - v_k = 0. \end{cases} \quad (2.26)$$

Re-substituting 2.26 into 2.24, the primal quadratic programming problem 2.23 becomes a dual form as follows

$$L(\mathbf{w}, b, \xi, \alpha, v) \quad (2.27)$$

$$\begin{aligned} &= \frac{1}{2} \sum_{k=1}^n \sum_{l=1}^n \alpha_k \alpha_l y_k y_l \varphi(\mathbf{x}_k)^T \varphi(\mathbf{x}_l) + \sum_{k=1}^n (c - \alpha_k - v_k) \xi_k \\ &- \sum_{k=1}^n \sum_{l=1}^n \alpha_k \alpha_l y_k y_l \varphi(\mathbf{x}_k)^T \varphi(\mathbf{x}_l) - \sum_{k=1}^n \alpha_k y_k b + \sum_{k=1}^n \alpha_k \\ &= -\frac{1}{2} \sum_{k=1}^n \sum_{l=1}^n \alpha_k \alpha_l y_k y_l \varphi(\mathbf{x}_k)^T \varphi(\mathbf{x}_l) + \sum_{k=1}^n \alpha_k. \end{aligned}$$

Finally, finding the optimal hyperplane in feature space  $Z$  is the solution to

$$\text{maximize } J_D(\alpha) = -\frac{1}{2} \sum_{k,l=1}^n y_k y_l \varphi(\mathbf{x}_k)^T \varphi(\mathbf{x}_l) \alpha_k \alpha_l + \sum_{k=1}^n \alpha_k, \quad (2.28)$$

$$\text{subject to } \begin{cases} \sum_{k=1}^n \alpha_k y_k = 0 \\ 0 \leq \alpha_k \leq c, k = 1, 2, \dots, n. \end{cases}$$

A key property of the SVM is that only the quantities that one needs to compute are scalar products, of the form  $\varphi(\mathbf{x}_k)^T \cdot \varphi(\mathbf{x}_l)$ . Therefore, it is convenient to introduce the so-called *kernel function*  $K$ , that is

$$K(\mathbf{x}_k, \mathbf{x}_l) = \varphi(\mathbf{x}_k)^T \varphi(\mathbf{x}_l). \quad (2.29)$$

The definition of kernel function that satisfies Mercer's theorem can be used as inner-product. Two examples of kernels used in SVM are

**Polynomials:**

$$K(\mathbf{x}_k, \mathbf{x}_l) = (\mathbf{x}_k^T \mathbf{x}_l + 1)^q, \text{ for } q > 0, \quad (2.30)$$

where  $q$  is a constant. When  $q = 1$ , the kernel is the linear kernel.

**Radial basis Function:**

$$K(\mathbf{x}_k, \mathbf{x}_l) = e^{-\left(\frac{\|\mathbf{x}_k - \mathbf{x}_l\|^2}{\sigma^2}\right)}, \quad (2.31)$$

where  $\sigma$  is a positive parameter to control the radius. Here we only show that the Gaussian (RBF) kernel indeed there is an inner product of two vectors in an infinite dimensional space. Assume  $\mathbf{x} \in R$  and  $\sigma > 0$

$$\begin{aligned} K(\mathbf{x}_k, \mathbf{x}_l) &= e^{-\left(\frac{\|\mathbf{x}_k - \mathbf{x}_l\|^2}{\sigma^2}\right)} = e^{-\left(\frac{(\mathbf{x}_k - \mathbf{x}_l)^2}{\sigma^2}\right)} \\ &= e^{-\frac{\mathbf{x}_k^2}{\sigma^2} + \frac{2\mathbf{x}_k\mathbf{x}_l}{\sigma^2} - \frac{\mathbf{x}_l^2}{\sigma^2}} \\ &= e^{-\frac{\mathbf{x}_k^2}{\sigma^2} - \frac{\mathbf{x}_l^2}{\sigma^2}} \left[1 + \left(\frac{2\mathbf{x}_k\mathbf{x}_l}{\sigma^2}\right) \cdot \frac{1}{1!} + \left(\frac{2\mathbf{x}_k\mathbf{x}_l}{\sigma^2}\right)^2 \cdot \frac{1}{2!} + \dots\right] \\ &= e^{-\frac{\mathbf{x}_k^2}{\sigma^2} - \frac{\mathbf{x}_l^2}{\sigma^2}} \left[1 \cdot 1 + \sqrt{\frac{2}{1! \cdot \sigma^2}} \mathbf{x}_k \cdot \sqrt{\frac{2}{1! \cdot \sigma^2}} \mathbf{x}_l + \sqrt{\frac{2^2}{2! \cdot (\sigma^2)^2}} \mathbf{x}_k^2 \cdot \sqrt{\frac{2^2}{2! \cdot (\sigma^2)^2}} \mathbf{x}_l^2 + \dots\right] \\ &= \varphi(\mathbf{x}_k)^T \varphi(\mathbf{x}_l), \end{aligned} \quad (2.32)$$

where

$$\varphi(\mathbf{x}) = e^{-\frac{\mathbf{x}^2}{\sigma^2}} \left[1, \sqrt{\frac{2}{1! \cdot \sigma^2}} \mathbf{x}, \sqrt{\frac{2^2}{2! \cdot (\sigma^2)^2}} \mathbf{x}^2, \dots\right]^T. \quad (2.33)$$

Finally, the nonlinear SVM classifier takes into the form

$$y(\mathbf{x}) = \text{sign}\left[\sum_{k=1}^n \alpha_k y_k K(\mathbf{x}_k, \mathbf{x}_l) + b\right]. \quad (2.34)$$

We can see that only support vector will affect the result in the prediction stage. In general, the number of support vector is not large. Therefore we can say SVM is used to find

important data (support vector) from training data.

### 2.3 Decision Tree

Decision tree has been constructed and used for data mining and classification, this technique is helpful to reveal explicit relationship between attributes among huge dataset. The decision tree is constructed in a recursive, top-down and divide-and-conquer manner. A decision tree consists of three types of nodes including decision nodes, chance nodes and end nodes. There are three popular rules applied into automatic creation of classification trees. The Gini rule splits off a single group as large as possible, whereas the entropy and twoing rules find multiple groups comprising as close to half the samples as possible. Both of the algorithms process recursively down the tree until stopping criteria.

The Gini rule is typically used by programs that induce decision trees using the CART algorithm. Gini rule is based on squared probabilities of membership for each target category in the node. It reaches its zero when all cases in the node fall into target category. Suppose  $y$  values are in  $\{1, 2, \dots, m\}$ , and let  $f(i, j)$  = probability of getting value  $j$  in node  $i$ . That is,  $f(i, j)$  is the proportion of records assigned to node  $i$  for which  $y = j$ .

$$I_G(i) = 1 - \sum_{j=1}^m f(i, j)^2 = \sum_{j \neq k} f(i, j)f(i, k) \quad (2.1)$$

Information gain is used by programs that are based on the ID3, C4.5 and C5.0 algorithm.

$$I_E(i) = - \sum_{j=1}^m f(i, j) \log_2 f(i, j) \quad (2.2)$$

Decision trees have several advantages. First, it is simple to understand and requires little data preparation. People can easy to understand decision tree models after a brief explanation and data is no need to normalization. Second, it is possible to validate a model by using statistical tests and perform well with large data with short time. Large amounts of data can be

analyzed using personal computers to enable stakeholders to take decisions based on its analysis.

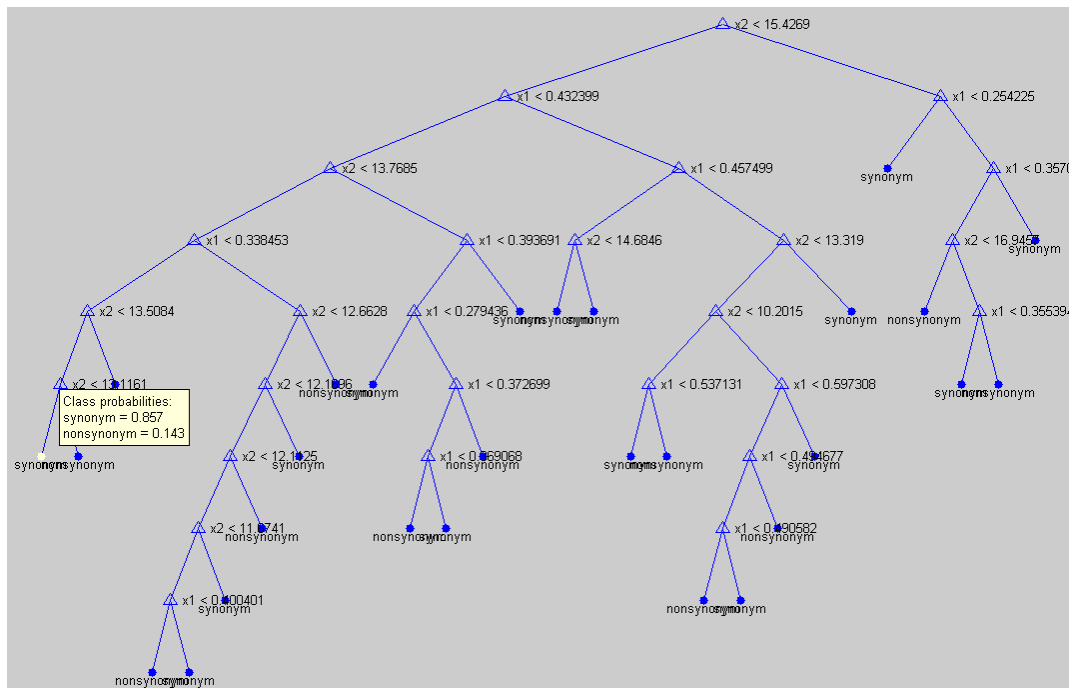


Figure 2.4: Decision tree

## 2.4 SNOMED-CT

SNOMED-CT stands for Systemized Nomenclature of Medicine Clinical Term is an ontological resource that has a wide coverage of the clinical domain. It is produced by the College of American Pathologists. SNOMED-CT is used for indexing clinical decision support, clinical trials, electronic medical records, ICU monitoring, medical research studies, computerized physician order entry, disease surveillance, imaging indexing and consumer health information services. The current version included in *UMLS* in May 2004 (2004AA) contains more than 360,000 concepts, 975,000 synonyms and 1,450,000 relationships organized into 18 hierarchies. The concepts and their descriptions are linked with semantic relationships including associated etiology, associated morphology, is-a, assists, treats, prevents, has property, has specimen, associated topography, has object, has manifestation, associated with, classifies, clinically associated with, has ingredient, mapped to, mapped from,

measures, used by, anatomic structure is physical part of.

## 2.5 MeSH

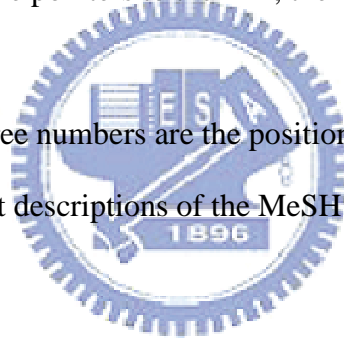
MeSH (Medical Subject Headings) [5, 6] is a hierarchical ontology of medical terminologies suggested by the U.S National Library of Medicine. There are 21,973 main headings and 15 subtrees in MeSH (22,568 in 2004). MeSH concepts correspond to MeSH describes terms of several property, the most important of them are the following:

**MeSH Headings (MH):** These are term names or identifiers used in MEDLINE as the indexing terms for documents. A MH term belongs to a concept, and is to label the meaning that corresponds to the concept reflects.

**Entry Terms:** These terms are pointers to the MH, there are the synonym terms of the MH with the same concept.

**MeSH Tree Number:** The tree numbers are the positions of the terms in the MeSH.

**MeSH Scope Note:** The text descriptions of the MeSH terms. This piece of text provides a type of definition.



## 2.6 Semantic Similarity Measurement Methods

This section introduces several ontology-based methods for computing the similarity between concepts or classes. Semantic similarity measures is useful for performing tasks such as retrieving results to user queries, representation and redundancy of retrieved resources, and checking ontology for coherency.

### 2.6.1 Edge-Counting Measures

The first category to measure semantic similarity considers where two concepts  $c1$  and  $c2$  are in the taxonomy. The following measurement based on a simplified version of spreading activation theory [8, 9]. The more similar two concepts are, the more links there

are between the concepts and the more closely related they are [10].

Wu and Palmer [11]: This similarity measure considers the position of concepts  $c_1$  and  $c_2$  related to the position of the lowest common concept  $c$ . As there may be multiple parents for each concept, two concepts can share parents by multiple paths.

$$sim_{W\&P}(c_1, c_2) = \frac{2H}{N_1 + N_2 + 2H} \quad (2.35)$$

$N_1$  and  $N_2$  is the number of edge from  $c_1$  and  $c_2$  respectively to the lowest common concept  $c$ , and  $H$  is the number of edge from  $c$  to the root of the taxonomy. It ranges from 1 to 0.

Li [12]: The following similarity measure, which combines the shortest path length between two concepts  $c_1$  and  $c_2$ ,  $L$ , and the depth in the taxonomy of the lowest common concept  $c$ ,  $H$ , in a non-linear function.

$$sim_{Li}(c_1, c_2) = e^{-\alpha L} \cdot \frac{e^{\beta H} - e^{-\beta H}}{e^{\beta H} + e^{-\beta H}} \quad (2.36)$$

where  $\alpha \geq 0$  and  $\beta \geq 0$  scaling the contribution of shortest path length and depth respectively. The optimal parameters of  $\alpha$  and  $\beta$  are 0.2 and 0.6 respectively. It is thus obvious that this measure ranges from 1 to 0.

## 2.6.2 Information Content Measures

The information content is estimated by the frequency of that concept in a large corpus of text. Information content requires the count of frequency of every concept include the frequency of all subsumed concepts in a hierarchy. For instance, the frequency for the concept of disease would include frequency of influenza and tuberculosis. The concept corresponds to the root of the hierarchy has the maximum frequency, so it includes the frequency of all other concept in the hierarchy. Thus, the frequency of the higher concepts in the hierarchy is always equal or greater than the lower concepts in the hierarchy in the hierarchy.



The information content of each concept  $c$  is computed as following:

$$IC(c) = -\log\left(\frac{freq(c)}{freq(root)}\right), \quad (2.37)$$

Resnik [13]: This measure uses the information content of the shared parents.

$$sim_{res}(c_1, c_2) = IC(lcs(c_1, c_2)), \quad (2.38)$$

If two terms share more information in common then the more similar they are. The information shared by two terms is indicated by the information content that subsume them in the hierarchy. This measure provides us with information such as the size of the corpus. A large numerical value indicates a large corpus. Furthermore, the score from comparing a term with itself depends on where in the hierarchy the term is. The less the term occurs the higher the score of the term.

Jiang and Conrath [14]: Scale the information content of the subsuming concept by the information content of the individual concepts. Jiang and Conrath are different. The Jiang and Conrath compute the inverse of similarity of concepts  $c1$  and  $c2$  as:

$$dist_{jen}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \cdot IC(lcs(c_1, c_2)) \quad (2.39)$$

### 2.6.3 Feature-Based Measures

Until now, the features of the terms are not taken into account. However, these features of a term contain valuable information about the term. The following measure including the features of terms in order to compute similarity between different concepts, but it ignores the position of the terms in the hierarchy.

Tversky [15]: This measure is based on the features of the terms. We suppose that each term is described by a set of words indicating its properties. If two terms have more common characteristics and the less non-common characteristics, the more similar the terms are.

$$sim_{Tversky}(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \kappa|C_1 \setminus C_2| + (\kappa - 1)|C_2 \setminus C_1|} \quad (2.40)$$

#### 2.6.4 Hybrid Measures

This method compare two concepts  $c1$  and  $c2$  combine some of the above approaches, considering the path connecting the two concepts in the hierarchy.

Rodriguez [16]: This approach can be used for single or cross ontology similarities. The similarity function is a weighted sum of the similarity values for features, neighborhoods and synonym sets.

$$S(a^p, b^q) = \omega_w \cdot S_w(a^p, b^q) + \omega_u \cdot S_u(a^p, b^q) + \omega_n \cdot S_n(a^p, b^q) \quad (2.41)$$

where  $w_w + w_u + w_n = 1$ ,  $S_w$ ,  $S_u$  and  $S_n$  are similarity functions. The functions  $S_w$ ,  $S_u$  and  $S_n$  are the similarity between synonym sets, features and neighborhoods of ontology p and b of ontology q and are calculated by equation 2.42.

$$S(a, b) = \frac{|A \cap B|}{|A \cap B| + \alpha|A \setminus B| + (1 - \alpha)|B \setminus A|} \quad (2.42)$$

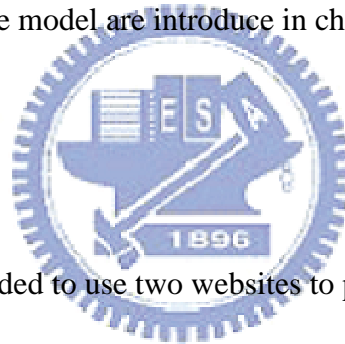
In this method  $\alpha$  is computed according to equation 2.43, but here  $\alpha$  is computed as a factor of the depth where the two compared concepts are in each hierarchy.

$$\alpha(c_1, c_2) = \begin{cases} \frac{d(c_1, c_{mis})}{d(c_1, c_2)}, & d(c_1, c_{mis}) \leq d(c_2, c_{mis}); \\ 1 - \frac{d(c_1, c_{mis})}{d(c_1, c_2)}, & d(c_1, c_{mis}) > d(c_2, c_{mis}). \end{cases} \quad (2.43)$$

where  $d(c1, c2) = d(c1, cmis) + d(c2, cmis)$ .

## Chapter 3 Methodology

We propose a method which exploits page-count to measure semantic similarity between a given pair of concepts. In chapter 3.1, we describe our sample construction. In chapter 3.2, we describe our feature definitions. We then describe a feature selection strategy in chapter 3.3. We rank the features by F-score according to their ability to express semantic similarity. We use two-class support vector machines (SVMs) and decision tree to find the optimal combination of features and training samples. The SVM and decision tree are trained to classify synonymous term-pairs and non-synonymous term-pairs and convert the output of SVM and decision tree into a posterior probability. We define the semantic similarity between two concepts as the posterior probability that they belong to the synonymous-terms (positive) class. The SVM and decision tree model are introduced in chapter 3.4 and chapter 3.5, respectively.



### 3.1 Sample Construction

For our experiment we decided to use two websites to provide synonymous and non-synonymous training sets from which our system to train a classifier. Our training set was drawn from the MedTerms Dictionary section of the website MedicineNet.com(<http://www.medterms.com/script/main/hp.asp>)(shown in figure 3.1). For the synonymous training set, we select one term from MedTerms Dictionary randomly and manually then query the website synonyms.net(<http://www.synonyms.net/synonym/>)(shown in figure 3.2) for the synonym. We repeat this procedure until 1500 synonymous term pairs was collected. For the non-synonymous training set, we select two terms from MedTerms Dictionary randomly and check synonyms.net(<http://www.synonyms.net/synonym/>) to make sure that the term pair was not synonymous. We repeat this procedure until 1500 non-synonymous term pairs was collected.

SEARCH

[About Us](#) | [Privacy Policy](#) | [Site Map](#)  
June 18, 2009

[Home](#)
[Picture Slideshows](#)
[Diseases & Conditions](#)
[Symptoms & Signs](#)
[Procedures & Tests](#)
[Medications](#)
[Health & Living](#)
[News & Views](#)
[MedTerms Dictionary](#)

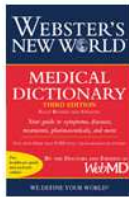
[home](#) > [medterms medical dictionary a-z list](#) - a

## MedTerms Medical Dictionary A-Z List - A

MedicineNet provides reliable [doctor pro](#) health and medical information.

MedicineNet Authored

Webster's New World  
Medical Dictionary  
[Learn more](#) »



MedTerms Medical Word

- [A](#)
- [B](#)
- [C](#)
- [D](#)
- [E](#)
- [F](#)
- [G](#)
- [H](#)
- [I](#)
- [J](#)
- [K](#)
- [L](#)
- [M](#)
- [N](#)
- [O](#)
- [P](#)

### Browse Centers

- [Allergies](#)
- [Alzheimer's](#)
- [Arthritis](#)
- [Asthma](#)
- [Blood Pressure](#)
- [Cancer](#)
- [Cholesterol](#)
- [Chronic Pain](#)
- [Cold & Flu](#)
- [Depression](#)
- [Diabetes](#)
- [Dictionary](#)
- [Digestion](#)

- [Pregnancy](#)
- [Senior Health](#)
- [Sexual Health](#)
- [Skin](#)
- [Sleep](#)
- [Thyroid](#)
- [Travel Health](#)
- [Women's Health](#)
- [650+ More Topics](#)

Search MedicineNet OR select from the options above.

 SEARCH

A → [Aa-AbAc-AcAd-AdAe-AgAh-ALAm-AmAn-AnAo-APAg-ArAs-AsAt-AtAu-AuAv-Az](#)

### Aa-Ab

- [A \(adenine\)](#)
- [A-](#)
- [a-](#)
- [A-T](#)
- [A. baumannii](#)
- [a.c.](#)
- [AIC](#)
- [AAA](#)
- [AAAAS](#)
- [AAD](#)
- [AAFP \(American Academy of Family Physicians\)](#)
- [AAMC \(Ass Am Medical Colleges\)](#)
- [AANAT](#)
- [AAO](#)
- [AAOS \(American Academy of Orthopaedic Surgeons\)](#)
- [AAP](#)
- [Aarskog syndrome](#)
- [Aarskog-Scott syndrome](#)
- [Aase-Smith syndrome I](#)
- [Aase-Smith syndrome II](#)

**Could Your Pain Be Fibromyalgia?**

ASSESS YOUR PAIN

**Important Safety Information**

Cymbalta® (duloxetine HCl) is approved for the treatment of depression and generalized anxiety disorder, and for the management of diabetic peripheral neuropathic pain and fibromyalgia.

**Cymbalta**  
duloxetine HCl  
DELAYED RELEASE CAPSULES

**Safety Information and Boxed Warning**  
**Prescribing Information**  
**Medication Guide**

Cymbalta is approved for the management of fibromyalgia.

WebMD

Angry. Sad. Unfocused.

**When Is It Depression?**

START HERE

**Could Your Pain Be Fibromyalgia?**

ASSESS YOUR PAIN

**Important Safety Information**

Figure 3.1: MedicineNet.com website



**What are Synonyms?**

Synonyms are different words with identical or at least similar meanings. Words that are synonyms are said to be *synonymous*, and the state of being a synonym is called *synonymy*. An example of synonyms are the words **car** and **automobile**, or **announcements** and **declarations**.

**What is Synonyms.net?**

Synonyms.net is the web's most comprehensive synonyms resource. To use Synonyms.net, simply type a word in the search box and click the **'Search'** button. A list of synonyms for the different word senses will be returned followed with images and translation options.

ADVERTISEMENT

[Online MBA](#)    [Wholesalers](#)

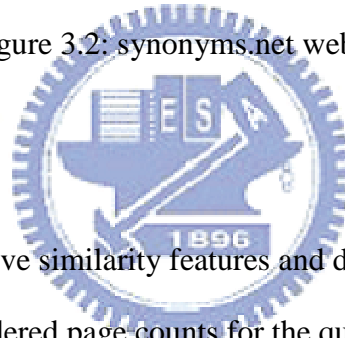
[Online MBA Degree](#)

[Addiction Treatment](#)

**Some sample synonyms**

Game	Amazing	Dictionary	Prominent
Feel	Infant	Upset	Moniker
Please	Insane	Resignation	Interesting
Melting	Good	Imagination	Trust
Stay	Funny	Believe	Modern
Sweet	Love	Manifest	Competition
War	Free	Surprise	Goal
Dog	Heaven	Peace	Joy

Figure 3.2: synonyms.net website



### 3.2 Feature Definitions

In this section we defined five similarity features and described ten lexico-syntactic pattern based features, we considered page counts for the query P AND Q as an approximation of co-occurrence of two concepts P and Q on the Web. However, page counts do not accurately express semantic similarity for the query P AND Q. For example, the search engine returns the page count 1150 for the query of abdomen AND breadbasket, whereas the same is 107000 for abdomen AND awareness. But, abdomen is more semantically similar to breadbasket than awareness, query for the page count of abdomen AND awareness is about one hundred times greater than those for the query abdomen and breadbasket. So we must consider the page counts not just for the query P AND Q, but also for the individual concepts P and Q to assess semantic similarity between P and Q.

We use five popular modified co-occurrence measures [17] Dice, Jaccard, Overlap (Simpson), PMI (Point-wise mutual information) and NGD (Normalized google distance) to

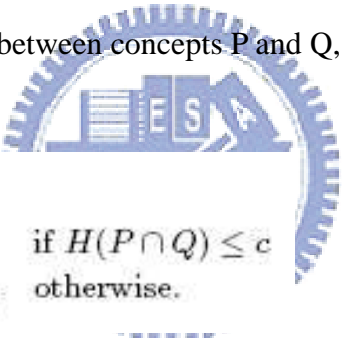
compute semantic similarity. For the remainder of this paper we use the notation  $H(P)$  to denote the page counts for the query  $P$ .

Therein,  $P \cap Q$  denotes the conjunction query  $P$  AND  $Q$ . It is possible that two concepts may appear on some pages purely accidentally given the scale and noise in Web data. In order to reduce the adverse effects attributable to random co-occurrences, if the page count for the query  $P \cap Q$  is less than a threshold  $c=5$  then we set the coefficient to zero.

WebDice coefficient is a variant of the Dice coefficient.  $\text{WebDice}(P, Q)$  is defined as follow:

$$\begin{aligned} & \text{WebDice}(P, Q) \\ &= \begin{cases} 0 & \text{if } H(P \cap Q) \leq c \\ \frac{2H(P \cap Q)}{H(P) + H(Q)} & \text{otherwise.} \end{cases} \end{aligned} \quad (3.1)$$

The WebJaccard coefficient between concepts  $P$  and  $Q$ ,  $\text{WebJaccard}(P, Q)$ , is defined as follow:



$$\begin{aligned} & \text{WebJaccard}(P, Q) \\ &= \begin{cases} 0 & \text{if } H(P \cap Q) \leq c \\ \frac{H(P \cap Q)}{H(P) + H(Q) - H(P \cap Q)} & \text{otherwise.} \end{cases} \end{aligned} \quad (3.2)$$

We define  $\text{WebOverlap}$ ,  $\text{WebOverlap}(P, Q)$ , as,

$$\begin{aligned} & \text{WebOverlap}(P, Q) \\ &= \begin{cases} 0 & \text{if } H(P \cap Q) \leq c \\ \frac{H(P \cap Q)}{\min(H(P), H(Q))} & \text{otherwise.} \end{cases} \end{aligned} \quad (3.3)$$

$\text{WebOverlap}$  is the modification to the  $\text{Overlap}$  (Simpson) coefficient.

We define  $\text{WebPMI}$  as a form of  $\text{PMI}$  using page counts as follow:

$$\begin{aligned} & \text{WebPMI}(P, Q) \\ &= \begin{cases} 0 & \text{if } H(P \cap Q) \leq c \\ \log_2 \left( \frac{H(P \cap Q) N}{H(P) H(Q)} \right) & \text{otherwise.} \end{cases} \end{aligned} \quad (3.4)$$

Here,  $N$  is the number of documents indexed by Google. Probabilities in equation 3.4 are estimated by the maximum likelihood principle. To calculate  $\text{PMI}$  accurately by equation 3.4,

we must know  $N$ , the number of documents indexed by Google. Although estimating the number of documents indexed by a search engine [18] is an interesting task, it is beyond the scope of this paper. We set  $N = 1000000000000$  according to the number of indexed pages reported by Google 7/25/2008 10:12:00 AM.

The following equation is developed by Rudi L. Cilibrasi and Paul M.B. Vitányi [19] which is based on information distance and Kolmogorov complexity using Google as search engine and the web as database. The method is applicable to other search engines and databases. We apply the equation as a feature to construct a method to automatically extract similarity of words and phrases from the web using Google page counts. The web is the largest database, and the context information entered by billions of users averages out to provide automatic semantics of useful quality.

$$\begin{aligned}
 & \text{NGD}(P, Q) \\
 &= \begin{cases} 0 & \text{if } H(P \cap Q) \leq c \\ \frac{\max(\log H(P), \log H(Q)) - \log H(P \cap Q)}{\log N - \min(\log H(P), \log H(Q))} & \text{otherwise.} \end{cases} \quad (3.5)
 \end{aligned}$$

Phrases such as *known as*, *is a*, *part of*, *is an example of* all indicate various semantic relations. Some of such phrases are useful for capturing synonymous relation. For example, apoptosis known as programmed cell death is a commonly used pattern in our daily life. From this example, we form the pattern  $P \text{ known as } Q$ , where we replace the two concepts *Apoptosis* and *Programmed cell death* by two wildcards  $P$  and  $Q$ . By the phrase known as we can conclude that  $P$  and  $Q$  are synonymous concepts. But, identifying the exact set of words that convey the semantic relationship between two concepts is remaining a challenging problem which requires deeper semantic analysis. However, such an analysis is not feasible considering the numerous ill-formed sentences. It is uncertain which patterns are useful for capturing synonymy. John McCrae and Nigel Collier [20] 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.

We use the eleven patterns mentioned in John McCrae and Nigel Collier’s paper and replace \* by empty string, then we define the ten patterns shown in table 3.1 as our features. There are two reasons why we replace \* by empty string. First reason is Google did not provide query for the regular expression. Another reason is that in John McCrae and Nigel Collier’s experiment many of the patterns were inflexible and matched very rarely so they simply allowed \* to match the empty string,

For each pair of concepts, we replace two wildcards P and Q of the patterns in table 3.1 by two concepts and query Google search engine for the page counts. If the concepts are synonymous there will be more page counts than that are non-synonymous. However, page counts do not accurately express semantic similarity for the query. For example, the search engine returns the page count 92 for the query of “apoptosis known as programmed cell death”, whereas the same is 34 for “dengue fever known as breakbone fever”. Since apoptosis and programmed cell death are synonymous concepts so does dengue fever and breakbone fever. But the page count of “apoptosis known as programmed cell death” is about three times greater than those for the query “dengue fever known as breakbone fever”. So we must consider the page counts not just for the query P known as Q, but also for the P AND Q to assess semantic similarity between P and Q. So we divide the page count of P known as Q by the page count of P AND Q. For the remaining ten patterns we use the equation 3.6 to assess semantic similarity between P and Q.

$$\text{WebPattern}(P,Q) = \begin{cases} 0 & \text{if } H(\text{Pattern}) \leq c \\ \frac{H(\text{Pattern})}{H(P \cap Q)} & \text{otherwise.} \end{cases} \quad (3.6)$$

Table 3.1: Lexico-syntactic patterns

**Pattern**

---



---

*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*

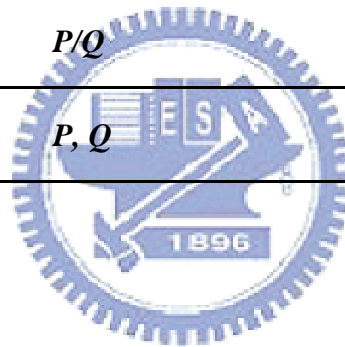
---

*P/Q*

---

*P, Q*

---



### 3.3 Feature Selection Strategy

The purpose of feature selection is to select some of the best features because data set contains features on the model often more than required for the establishment.

For example, the data set may contain 500 features to describe the characteristics of data set, but may only have 50 features will be used to create a specific model. If you are in the establishment of this model do not need those features so that we can reduce the need of CPU, memory and storage space.

Even if the resource is not a problem, you will usually want to remove unnecessary features, because they may reduce the quality of models have been exploring for the following reasons:

Certain feature is either cumbersome or superfluous. This situation will make it more difficult for meaningful patterns of information found.

To find the model of high-quality, most of the algorithm needs to provide high-dimension data sets much larger training data sets. Feature selection help to solve too many low-value information or high-value information on the problem of too few.

Generally speaking, the selection of features is to calculate the scores of each feature, and then only with the best scores of selected features. You can adjust the high threshold. Feature selection will be in shape before the implementation of the model, can automatically choose from the data sets are most likely to be used in the model features.

In this paper, we use F-score [21] our feature selection strategy. It is a simple technique which measures the discrimination of two sets of real numbers. Given training vectors  $x_k$ ,  $k = 1, \dots, n$ , if the number of positive and negative instances are  $n_+$  and  $n_-$  respectively, then the score of the  $i$ th feature is defined as:

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}, \quad (3.7)$$

where  $\bar{x}_i$ ,  $\bar{x}_i^{(+)}$ ,  $\bar{x}_i^{(-)}$  are the average of the  $i$ th feature of all, positive, and negative data sets, respectively;  $x_{k,i}^{(+)}$  is the  $i$ th feature of the  $k$ th positive instance, and  $x_{k,i}^{(-)}$  is the  $i$ th feature of the  $k$ th 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, we use this score as a feature selection strategy. We calculate the F-score with each features from 100 to 1500 training samples and averaged the scores for each features.

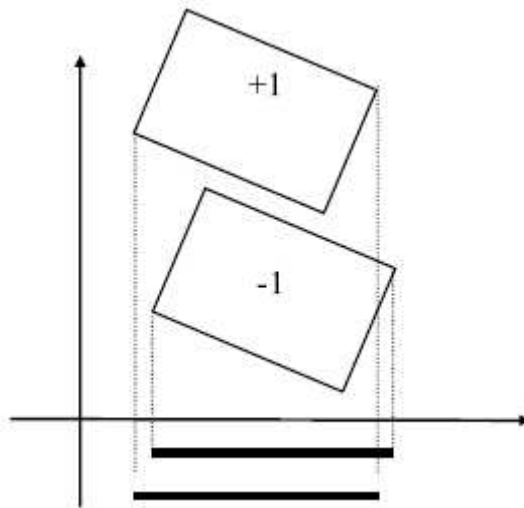


Figure 3.3: Both features of this data have low scores as in equation 3.6 the denominator is much larger than the numerator

### 3.4 Support Vector Machine Model

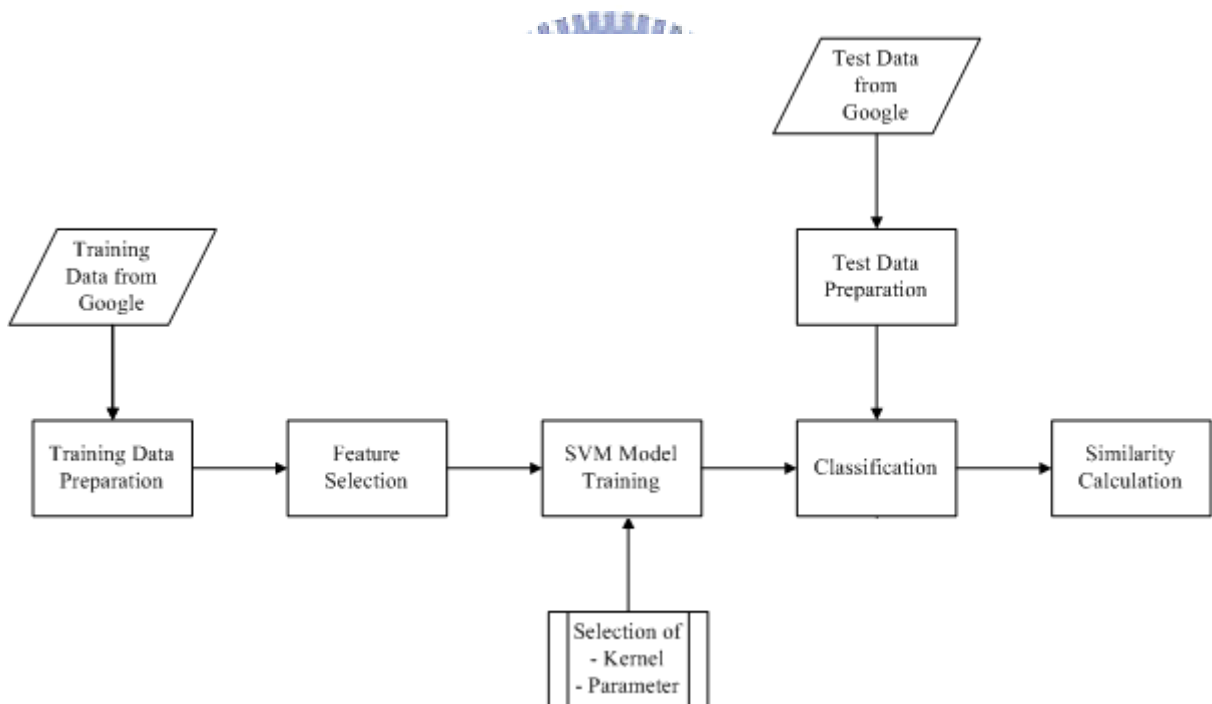


Figure 3.4: Support vector machine model flow chart

In section 3.2 we defined fifteen similarity scores using page counts. Section 3.3 described a strategy to rank the features according to their ability to express synonymy. In this

section we describe leverage of a semantic similarity measurement through integration of all the similarity scores described in previous sections.

For each pair of concepts  $(P, Q)$ , we create a feature vector  $F$ . First, we query Google and collect page counts for P, Q, P AND Q and ten lexico-syntactic patterns. Second, we calculate fifteen features by the equations mentioned in section 3.2. After that we use equation 3.6 to rank the features according to their ability to express synonymy. Finally we yields a 15 dimensional feature vector  $F$ . We form such feature vectors for all synonymous pairs (positive training samples) as well as non-synonymous pairs (negative training samples). We then train a two-class SVM with feature vectors. After we have trained a SVM using synonymous and non-synonymous pairs, we can use it to compute the semantic similarity between two given concepts. Following the same method we used to generate feature vectors for training, we create a feature vector  $F_0$  for the given pair of concepts  $(P_0, Q_0)$ , between which we need to measure the semantic similarity. The semantic similarity between  $P_0$  and  $Q_0$  as the posterior probability  $\text{Prob}(F_0 \setminus \text{synonymous})$  that feature vector  $F_0$  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. We use sigmoid functions to convert this distance into a posterior probability (see [22] for a detailed discussion on this topic). In our research we use libsvm 2.89 [23] toolbox including C-SVC ( $C=1$ ) and nu-SVC ( $\text{nu}=0.5$ ) to do the experiment (see [24] for a detailed discussion on this topic of C-SVC and nu-SVC).

### 3.5 Decision Tree Model

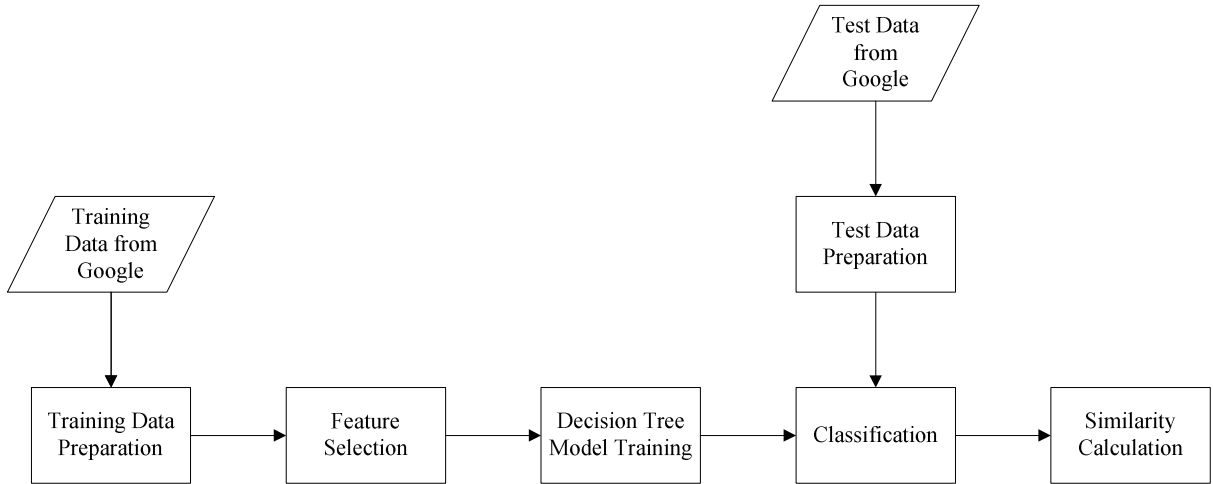


Figure 3.5: Decision tree model flow chart

In section 3.2 we defined fifteen similarity scores using page counts. Section 3.3 described a strategy to rank the features according to their ability to express synonymy. In this section we describe leverage of a semantic similarity measurement through integration of all the similarity scores described in previous sections.

For each pair of concepts  $(P, Q)$ , we create a feature vector  $F$ . First, we query Google and collect page counts for  $P$ ,  $Q$ ,  $P$  AND  $Q$  and ten lexico- syntactic patterns. Second, we calculate fifteen features by the equations mentioned in section 3.2. After that we use equation 3.6 to rank the features according to their ability to express synonymy. Finally we yields a 15 dimensional feature vector  $F$ . We form such feature vectors for all synonymous pairs (positive training samples) as well as non-synonymous pairs (negative training samples). We then train a two-class CART decision tree with feature vectors. After we have trained a CART decision tree using synonymous and non-synonymous pairs, we can use it to compute the semantic similarity between two given concepts. Following the same method we used to generate feature vectors for training, we create a feature vector  $F_0$  for the given pair of concepts  $(P_0, Q_0)$ , between which we need to measure the semantic similarity. The semantic similarity between  $P_0$  and  $Q_0$  as the posterior probability  $\text{Prob}(F_0 \setminus \text{synonymous})$  that feature vector  $F_0$

belongs to the synonymous (positive) class. In our research we use decision tree toolbox [25] in Matlab to do the experiment.



## Chapter 4 Experiment Results

There are several methodologies to assess the accuracy of similarity values computed by a given similarity measure [26]. One of them is to use the similarity measure in an application that requires similarity between concepts like information retrieval. Another is to compare the computed similarity scores of the measure against the human similarity scores using, for example, correlation coefficient (Pearson). Another methodology requires a dataset of concept pairs scored for similarity by experts. In our research, we calculate the correlation coefficient to evaluate the proposed measure.

We introduce two datasets of our experiments to evaluate the proposed semantic similarity measure in chapter 4.1. Then we introduce our experiment environment in chapter 4.2. After that we compare the similarity scores produced by the proposed measure against [27] dataset. We analyze the behavior of the proposed measure with the different number of features from 2 to 15, training samples from 100 to 1500 and classifiers including C-SVC, nu-SVC, based on four kernels (linear kernel SVM, SVM-2 (Polynomial kernel degree 2), SVM-3 (Polynomial kernel degree 3), and RBF), and decision tree in chapter 4.3. The correlations against [28] dataset are shown in chapter 4.4. The comparisons with other methods are shown in chapter 4.5.

### 4.1 Datasets

There are no standard human rating benchmark datasets in biomedical domain. To evaluate our methods, we used dataset 1 [27] contains 36 biomedical (MeSH) concept pairs. The human scores in this dataset are the average evaluated scores of reliable doctors. Table 4.1 contains the first 36 pairs of this dataset. The concept pairs in bold, in Table 4.1, are the ones that contains a term that was not found in SNOMED-CT.

The dataset 2 [28] of 30 concept pairs from Pedersen et al., which was annotated by 9

medical index experts and 3 physicians. The concept pairs in bold, in Table 4.2, are the ones that contains a term that was not found in MeSH. Each pair was annotated on a 4 scale: *unrelated*(1), *marginally related*(2), *related*(3) and *practically synonymous*(4). Table 4.2 contains only 30 pairs of this dataset. The average correlation between experts is 0.78, and between physicians is 0.68.

Table 4.1: Dataset 1 of 36 biomedical concept pairs

Concept 1	Concept 2	H
Anemia	Appendicitis	0.031
Dementia	Atopic Dermatitis	0.062
Bacterial Pneumonia	Malaria	0.156
Osteoporosis	Patent Ductus Arteriosus	0.156
Amino Acid Sequence	Anti-Bacterial Agents	0.156
Acquired Immunodeficiency Syndrome	Congenital Heart Defects	0.062
Otitis Media	Infantile Colic	0.156
Meningitis	Tricuspid Atresia	0.031
Sinusitis	Mental Retardation	0.031
Hypertension	Kidney Failure	0.5
Hyperlipidemia	Hyperkalemia	0.156
Hypothyroidism	Hyperthyroidism	0.406
Sarcoidosis	Tuberculosis	0.406
Vaccines	Immunity	0.593
Asthma	Pneumonia	0.375
Diabetic Nephropathy	Diabetes Mellitus	0.5
Lactose Intolerance	Irritable Bowel Syndrome	0.468
Urinary Tract Infection	Pyelonephritis	0.656
Neonatal Jaundice	Sepsis	0.187
Sickle Cell Anemia	Iron Deficiency Anemia	0.437
<b>Psychology</b>	<b>Cognitive Science</b>	0.593
Adenovirus	Rotavirus	0.437
Migraine	Headache	0.718
Myocardial Ischemia	Myocardial Infarction	0.75
Hepatitis B	Hepatitis C	0.562



Carcinoma	Neoplasm	0.75
Pulmonary Valve Stenosis	Aortic Valve Stenosis	0.531
Failure To Thrive	Malnutrition	0.625
Breast Feeding	Lactation	0.843
<b>Antibiotics</b>	<b>Antibacterial Agents</b>	0.937
Seizures	Convulsions	0.843
Pain	Ache	0.875
Malnutrition	Nutritional Deficiency	0.875
Measles	Rubeola	0.906
Chicken Pox	Varicella	0.968
Down Syndrome	Trisomy 21	0.875

Table 4.2: Dataset 2 of 30 biomedical concept pairs sorted in the order of the averaged physician's scores

Concept 1	Concept 2	Phy	Exp
Renal Failure	Kidney Failure	4	4
Heart	Myocardium	3.3	3
Stroke	Infarct	3	2.8
Abortion	Miscarriage	3	3.3
Delusion	Schizophrenia	3	2.2
Congestive Heart Failure	Pulmonary Edema	3	1.4
Metastasis	Adenocarcinoma	2.7	1.8
Calcification	Stenosis	2.7	2
<b>Diarrhea</b>	<b>Stomach Cramps</b>	2.3	1.3
Mitral Stenosis	Atrial Fibrillation	2.3	1.3
<b>Chronic Obstructive Pulmonary Disease</b>	<b>Lung Infiltrates</b>	2.3	1.8
Rheumatoid Arthritis	Lupus	2	1.1
Brain Tumor	Intracranial Hemorrhage	2	1.3
Carpal Tunnel Syndrome	Osteoarthritis	2	1.1
Diabetes Mellitus	Hypertension	2	1
Acne	Syringe	2	1
Antibiotic	Allergy	1.7	1.2
Cortisone	Total Knee Replacement	1.7	1
<b>Pulmonary Embolus</b>	<b>Myocardial Infarction</b>	1.7	1.2
Pulmonary Fibrosis	Lung Cancer	1.7	1.4

Cholangiocarcinoma	Colonoscopy	1.3	1
Lymphoid Hyperplasia	Laryngeal Cancer	1.3	1
Multiple Sclerosis	Psychosis	1	1
Appendicitis	Osteoporosis	1	1
<b>Rectal Polyp</b>	<b>Aorta</b>	1	1
Xerostomia	Alcoholic Cirrhosis	1	1
Peptic Ulcer Disease	Myopia	1	1
Depression	Cellulitis	1	1
<b>Varicose Vein</b>	<b>Entire Knee Meniscus</b>	1	1
Hyperlipidemia	Metastasis	1	1

## 4.2 Experiment Environment

- Hardware: CPU Intel Pentium 4, RAM 2.0GB.
- Software: Windows XP Professional, Matlab 7.3.0, LIBSVM 2.89.



## 4.3 Parameter Optimization

### 4.3.1 Classifier Models

We use C-SVC, nu-SVC, based on four kernels (linear kernel SVM, SVM-2 (Polynomial kernel degree 2), SVM-3 (Polynomial kernel degree 3), and RBF), and decision tree.

### 4.3.2 Number of features and training samples

In this section, we list the ranked features by our feature selection strategy (illustrated in Section 3.3). We use the following feature selection equation F-score. It is a function to measure the discrimination of two sets of real numbers. Results of the ranked features are shown in Table 4.3. Features with the highest  $F(i)$  value is NGD (rank=1,  $F(i)=0.2751$ ). Followed by a series of features such as WebPMI (rank=2,  $F(i)=0.237$ ),  $P(Q)$  (rank=3,  $F(i)=0.1648$ ) and  $P/Q$  (rank=2,  $F(i)=0.1632$ )

In the first experiment, in order to determine the optimum combination of features and

training samples, we trained the classifiers mentioned in 4.3.1 with 15 features (ranked according to their ability to capture the synonyms) and different numbers of samples starting from 100 to 1500 and calculated the correlation coefficient against the dataset 1, respectively.

Table 4.3: Features with highest F-scores

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

Experimental results using C-SVC with linear kernel are summarized in Figure 4.1. The maximum correlation coefficient of 0.758 is achieved with 9 features and 1500 training samples.

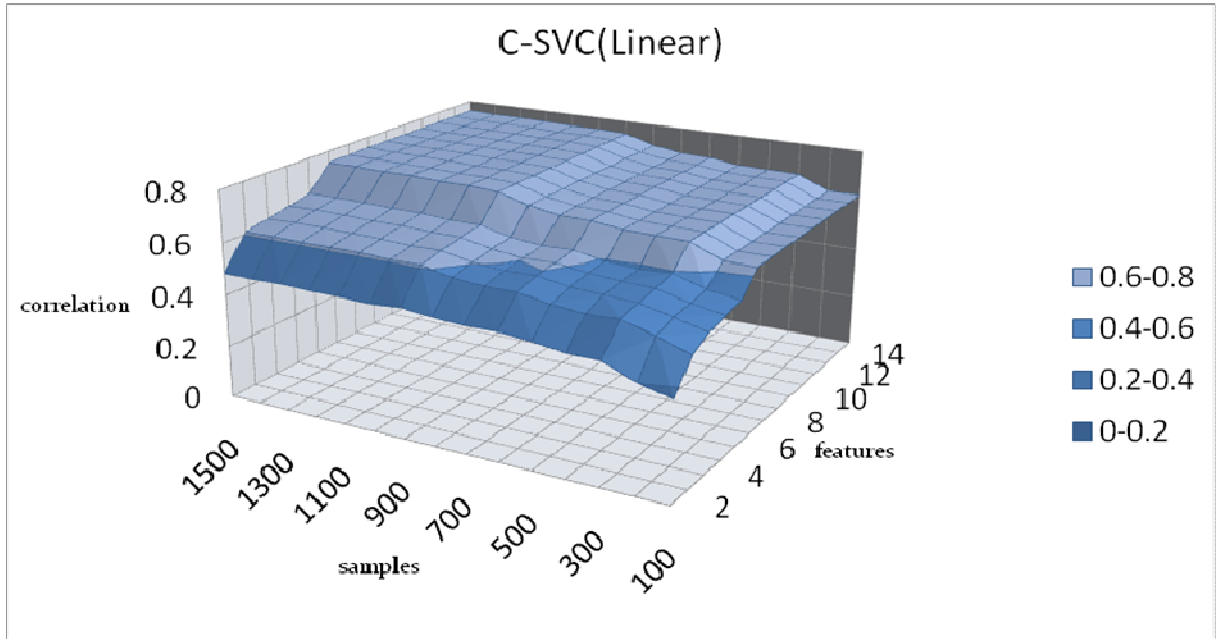


Figure 4.1: Correlation vs. No of features and training samples using C-SVC with linear kernel

Experimental results using C-SVC with polynomial degree=2 kernel are summarized in Figure 4.2. The maximum correlation coefficient of 0.776 is achieved with 7 features and 1200 training samples.

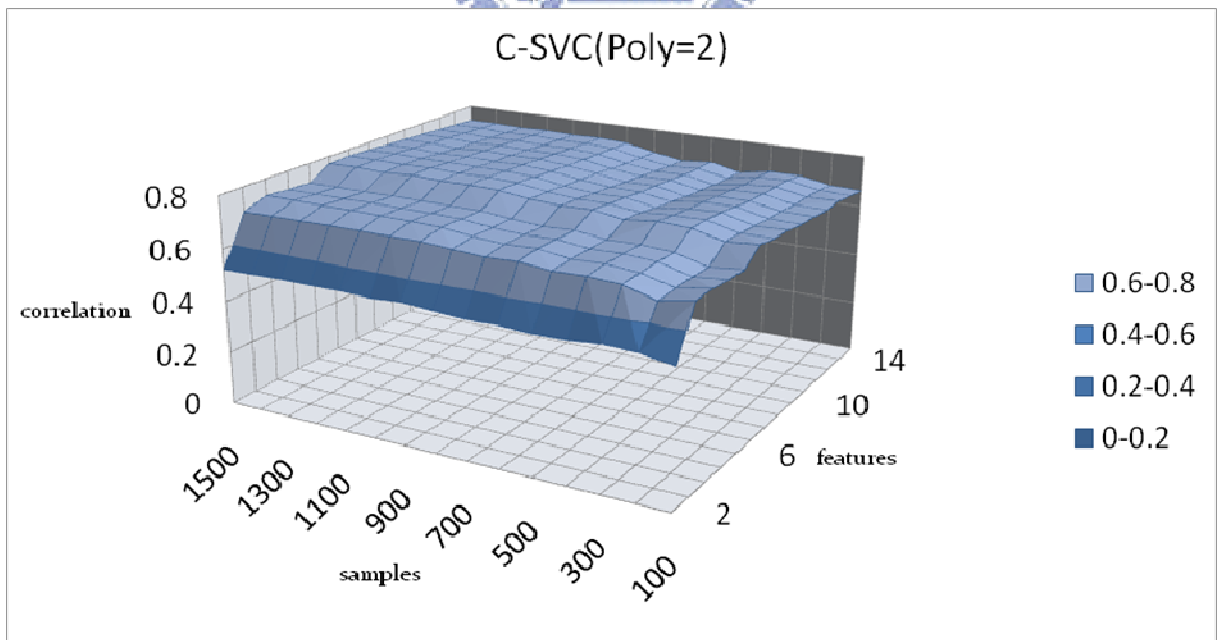


Figure 4.2: Correlation vs. No of features and training samples using C-SVC with polynomial degree=2 kernel

Experimental results using C-SVC with polynomial degree=3 kernel are summarized in Figure 4.3. The maximum correlation coefficient of 0.759 is achieved with 13 features and 300 training samples.

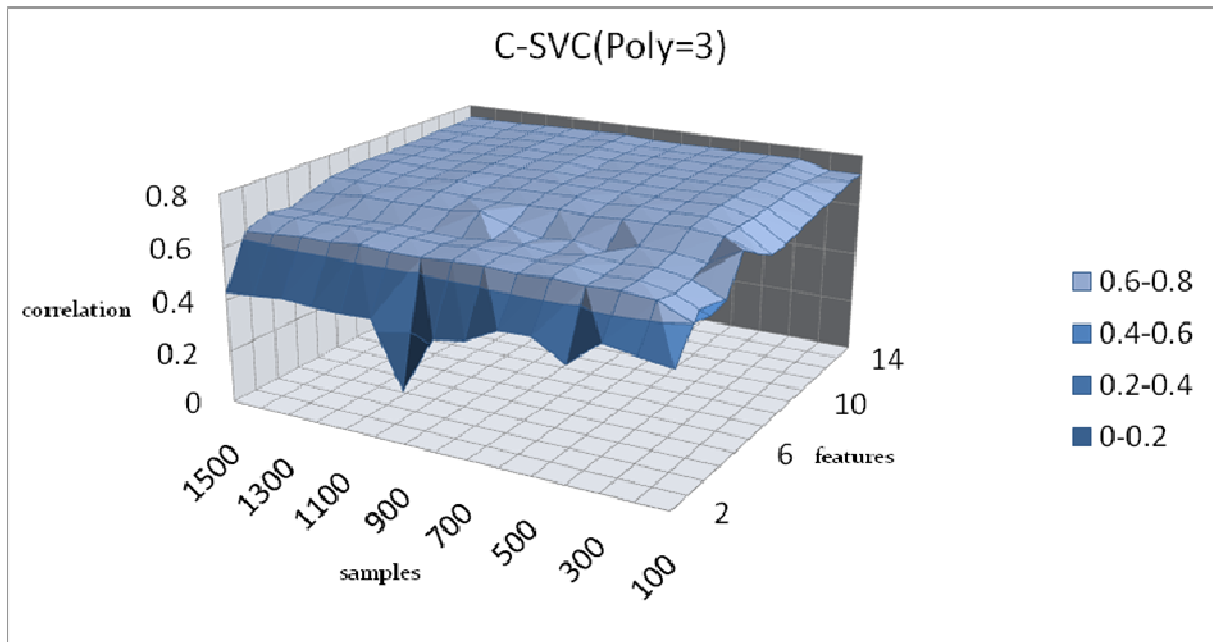


Figure 4.3: Correlation vs. No of features and training samples using C-SVC with polynomial degree=3 kernel

Experimental results using C-SVC with RBF kernel are summarized in Figure 4.4. The maximum correlation coefficient of 0.612 is achieved with 10 features and 1100 training samples.

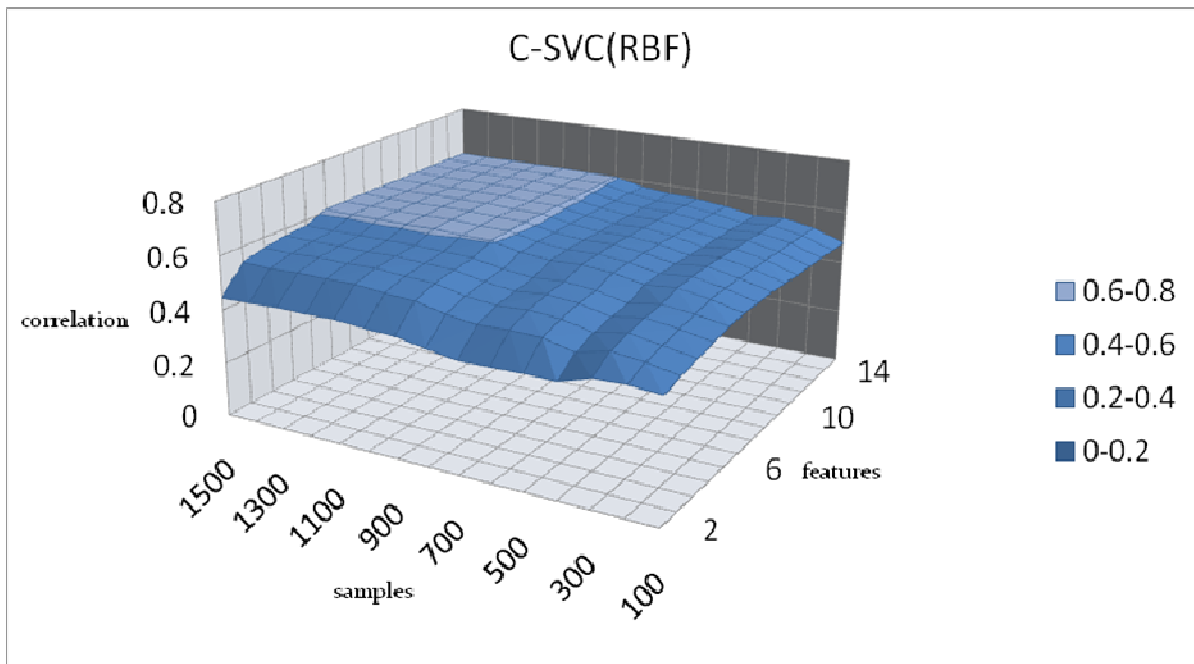


Figure 4.4: Correlation vs. No of features and training samples using C-SVC with RBF kernel

Experimental results using nu-SVC with linear kernel are summarized in Figure 4.5. The maximum correlation coefficient of 0.798 is achieved with 7 features and 900 training samples.

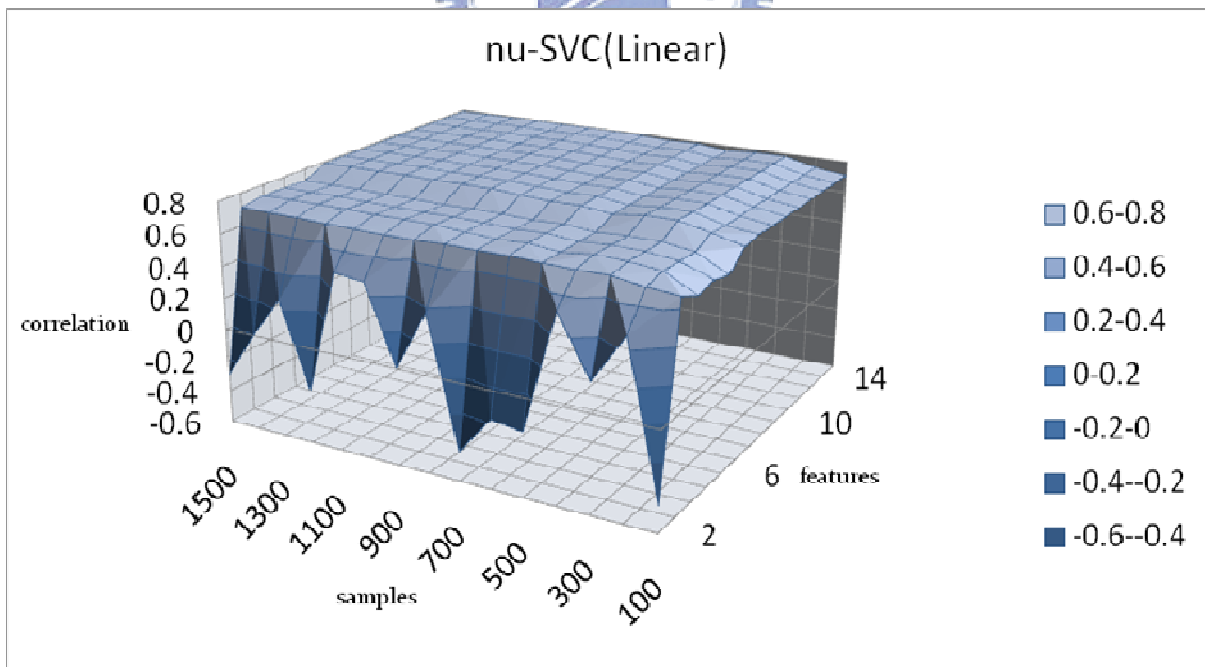


Figure 4.5: Correlation vs. No of features and training samples using nu-SVC with linear kernel

Experimental results using nu-SVC with polynomial degree=2 kernel are summarized

in Figure 4.6. The maximum correlation coefficient of 0.766 is achieved with 11 features and 300 training samples.

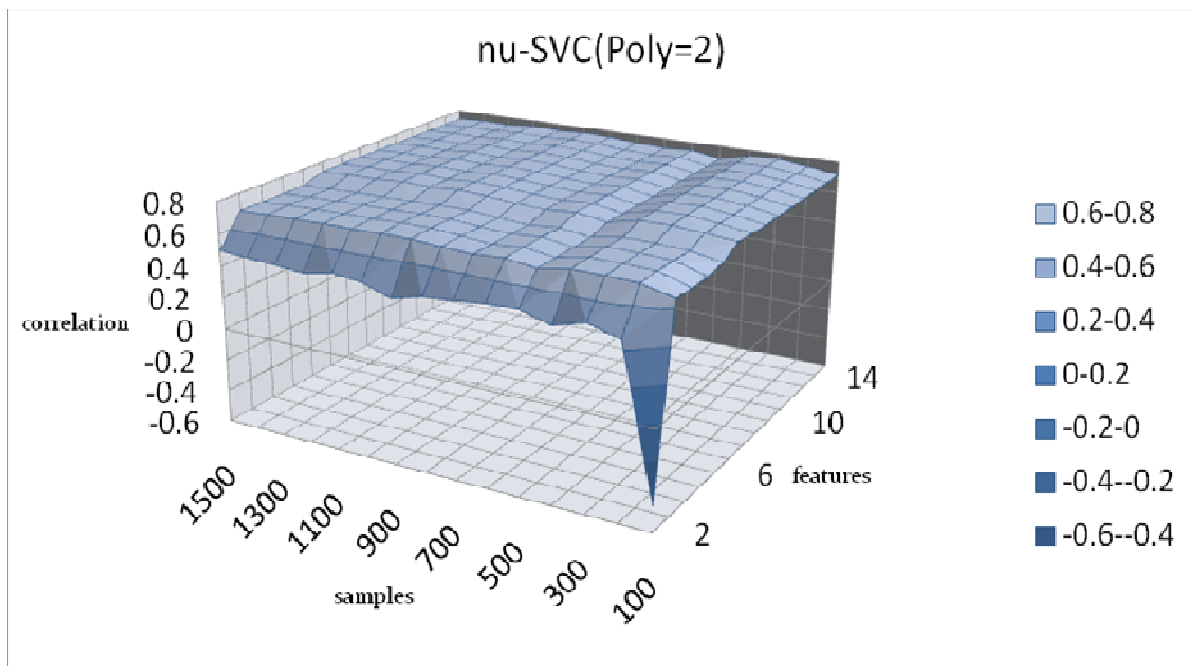


Figure 4.6: Correlation vs. No. of features and training samples using nu-SVC with polynomial degree=2 kernel

Experimental results using nu-SVC with polynomial degree=3 kernel are summarized in Figure 4.7. The maximum correlation coefficient of 0.736 is achieved with 12 features and 300 training samples.

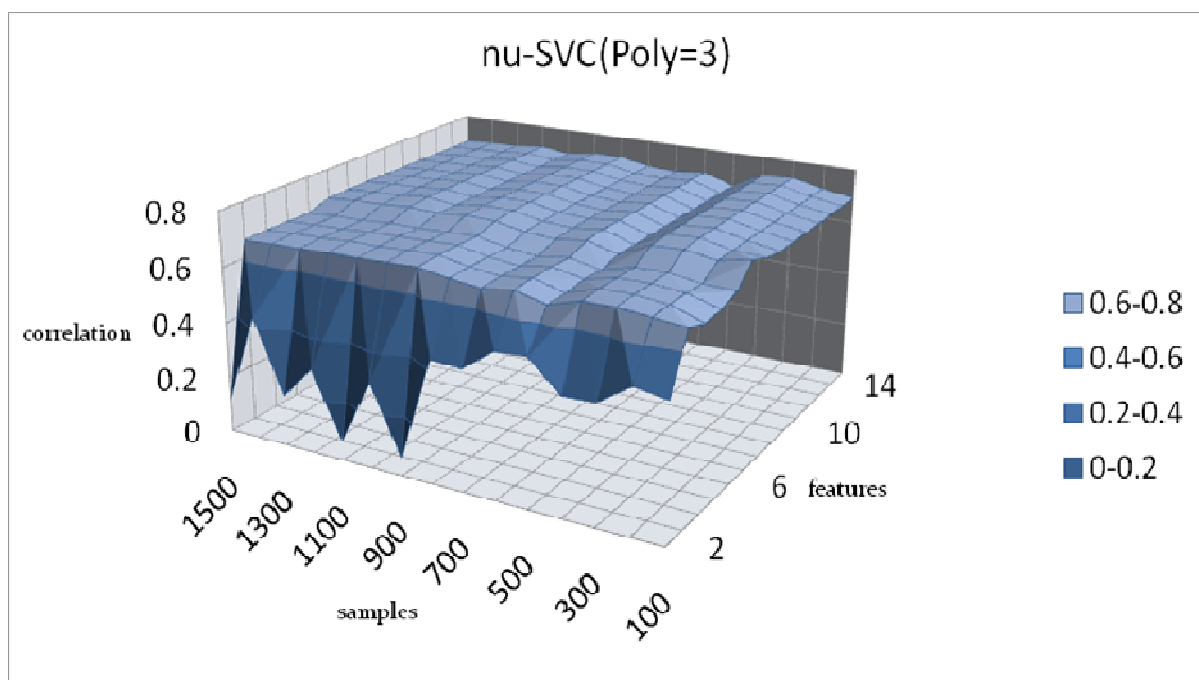


Figure 4.7: Correlation vs. No of features and training samples using nu-SVC with polynomial degree=3 kernel

Experimental results using nu-SVC with RBF kernel are summarized in Figure 4.8. The maximum correlation coefficient of 0.743 is achieved with 11 features and 100 training samples.

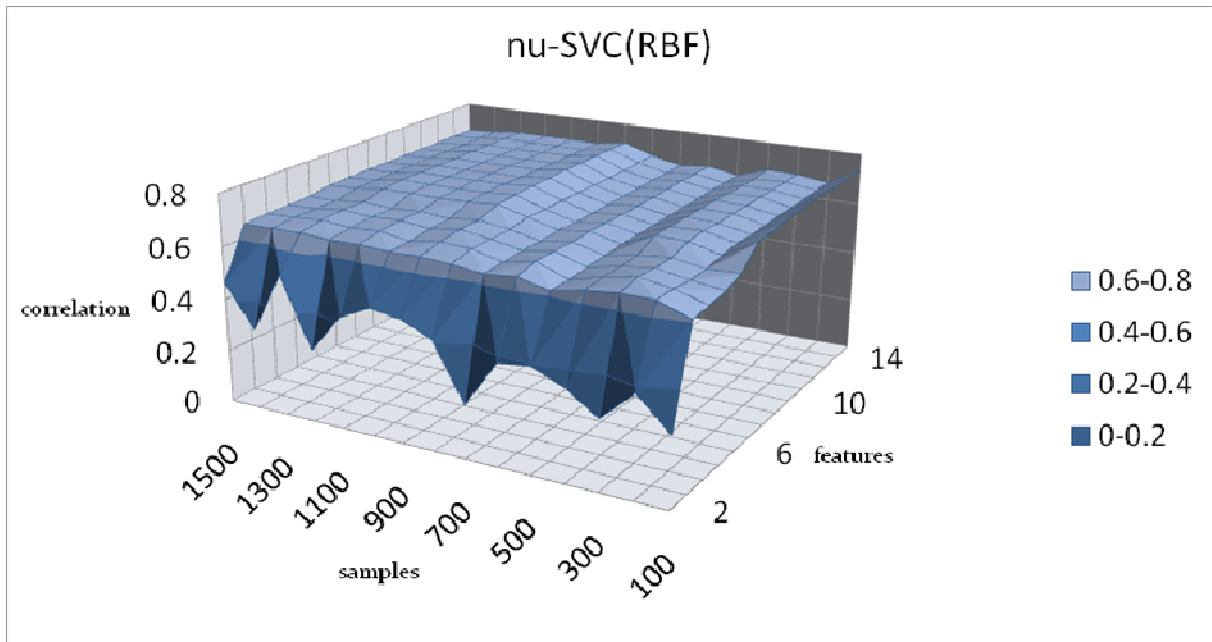


Figure 4.8: Correlation vs. No of features and training samples using nu-SVC with RBF kernel

Experimental results using decision tree are summarized in Figure 4.9. The maximum correlation coefficient of 0.734 is achieved with 5 features and 1300 training samples.



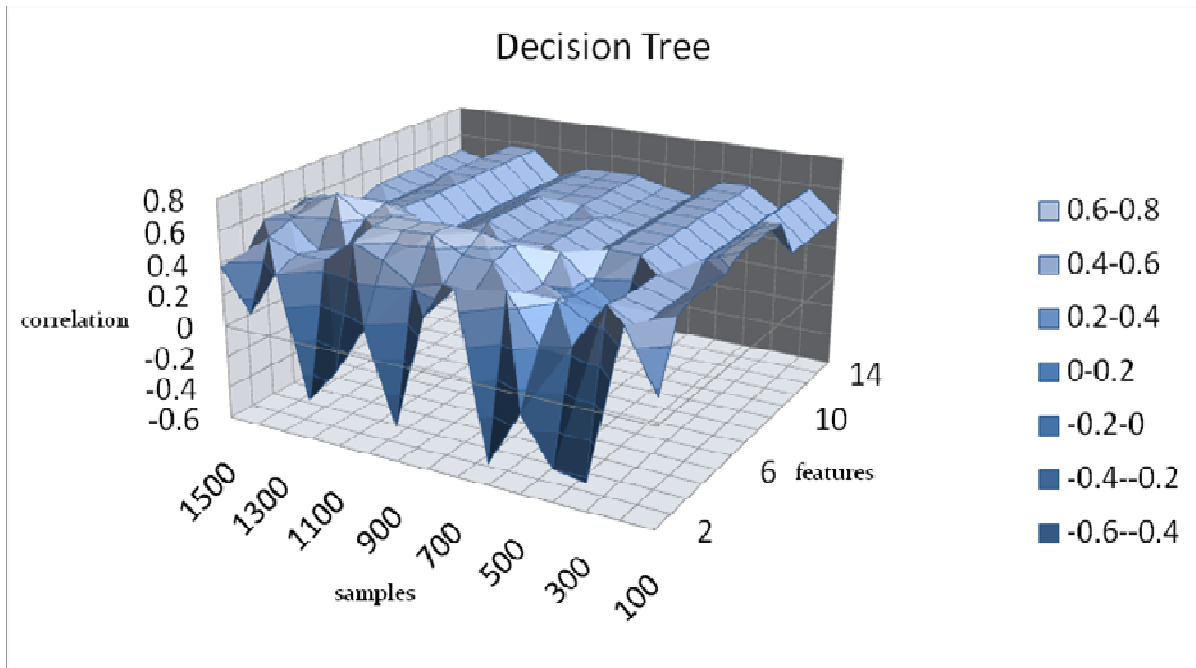


Figure 4.9: Correlation vs. No of features and training samples using decision tree

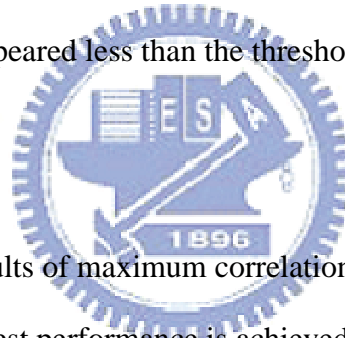
We experimented with different kernel types as shown in Table 4.4. Best performance is achieved with the linear kernel of nu-SVC, which reports a correlation coefficient of 0.798. When higher degree kernels such as quadratic(Polynomial degree=2) and cubic(Polynomial degree=3) of nu-SVC are used, correlation with the human ratings decreases. Second best is the C-SVC with quadratic(Polynomial degree=2) kernel, which reports a correlation coefficient of 0.776.

Table 4.4: Correlation vs. No 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)	<b>0.798</b>	<b>900</b>	<b>7</b>
nu-SVC(Poly=2)	0.766	300	11
nu-SVC(Poly=3)	0.736	300	12
nu-SVC(RBF)	0.743	100	11
Decision Tree	0.734	1300	5

In the second experiment, we trained the classifiers mentioned in 4.3.1 with the optimized feature numbers and sample numbers determined in the first experiment and calculated the correlation coefficient against the dataset 2 with 28 concept pairs out of 30. Because the concept *lung infiltrates* was not found in the SNOMEDCT terminology and the concept *entire knee meniscus* appeared less than the threshold  $c=5$  that we set in section 3.2.



#### 4.4 Results

Figure 4.10 shows the results of maximum correlation in dataset 1 of different classifiers mentioned in 4.3.1. Best performance is achieved with the linear kernel of nu-SVC, which reports a correlation coefficient of 0.798.

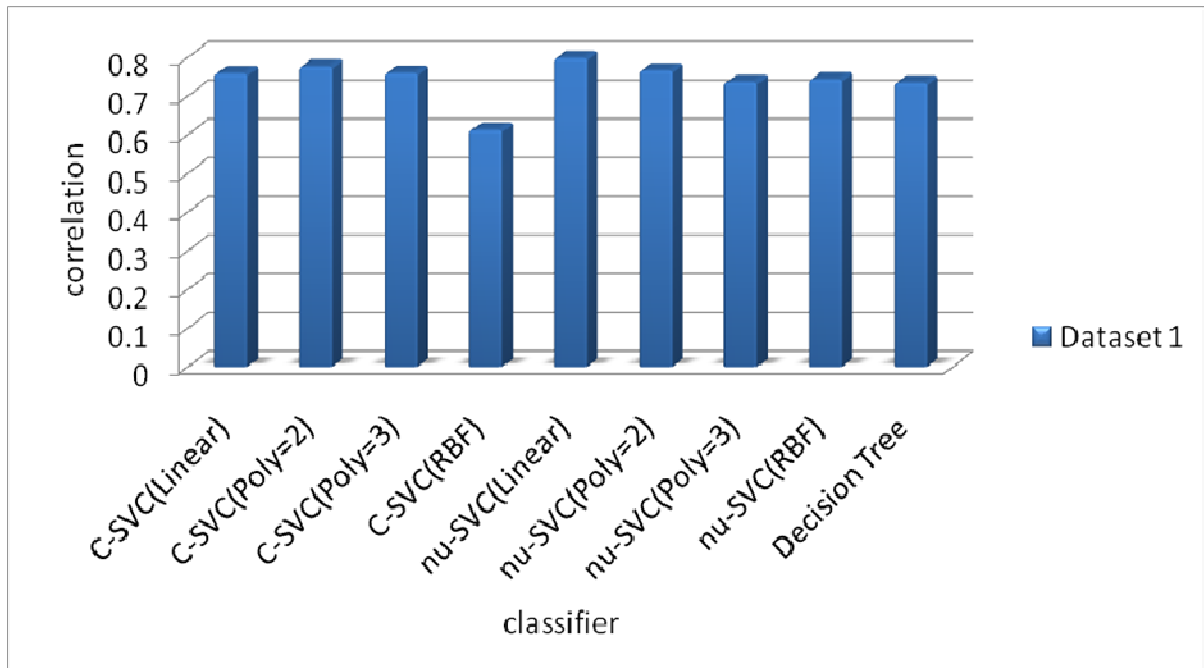


Figure 4.10: Correlation vs. Classifiers of dataset 1 with human scores

Figure 4.11 shows the results of maximum correlation in dataset 2 with physician scores of different classifiers mentioned in 4.3.1 and the optimized feature numbers and sample numbers determined in the first experiment. Best performance is achieved with the linear kernel of nu-SVC, which reports a correlation coefficient of 0.705.

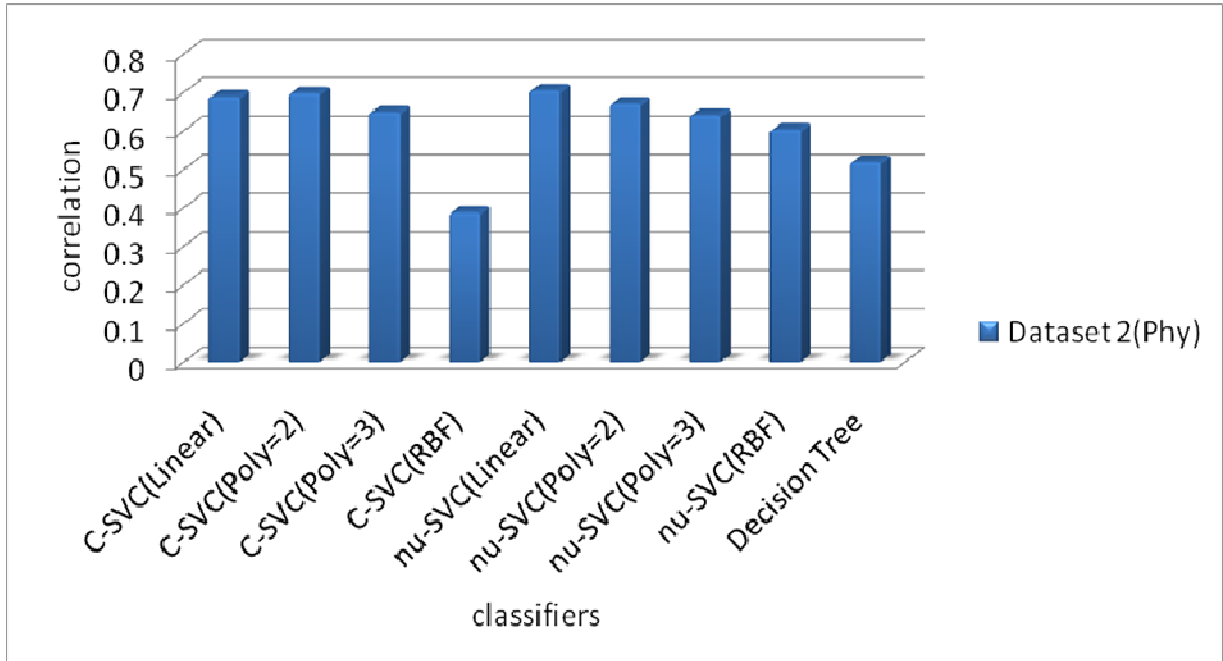


Figure 4.11: Correlation vs. Classifiers of dataset 2 with physician scores

Figure 4.12 shows the results of maximum correlation in dataset 2 with expert scores of different classifiers mentioned in 4.3.1 and the optimized feature numbers and sample numbers determined in the first experiment. Best performance is achieved with the linear kernel of nu-SVC, which reports a correlation coefficient of 0.496.

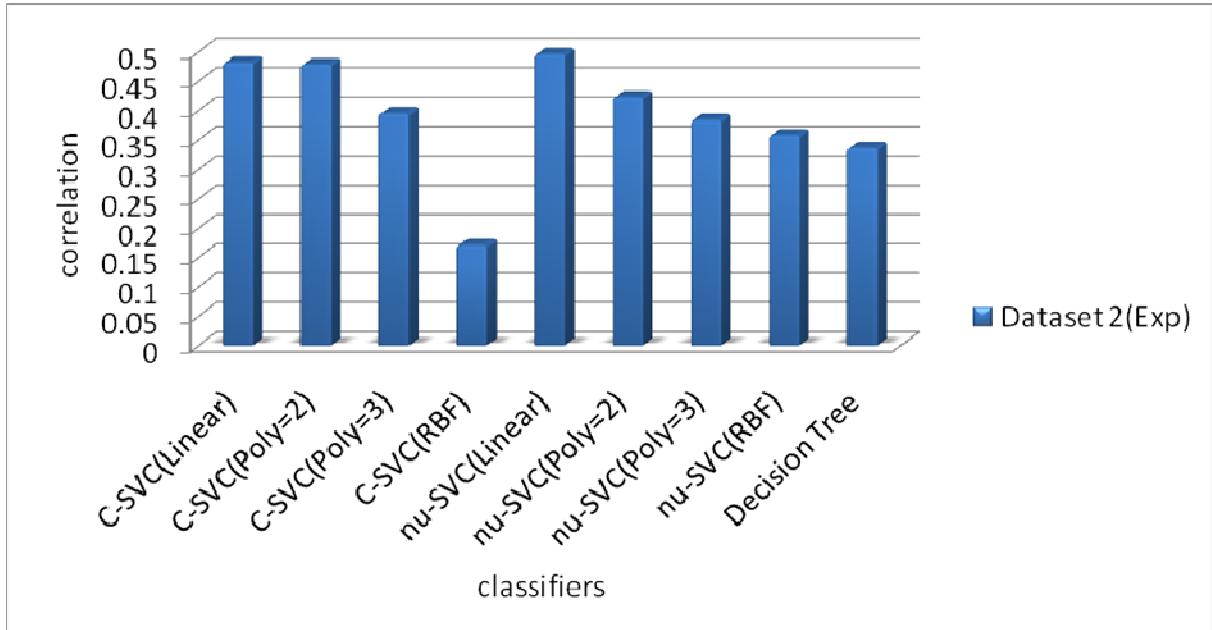


Figure 4.12: Correlations vs. Classifiers of dataset 2 with expert scores

Table 4.5 show that proposed method earns the highest correlation of 0.798 in dataset 1, 0.705 in dataset 2 with physician scores and 0.496 in dataset 2 with expert scores using C-SVC with linear kernel.

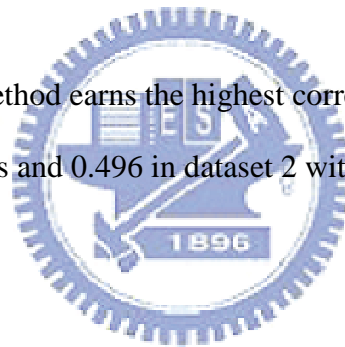


Table 4.5: Correlation vs. Dataset 1 and Dataset 2 with physician scores and expert scores of different models

Model	Dataset 1	Dataset 2(Phy)	Dataset 2(Exp)
C-SVC(Linear)	0.758	0.689	0.482
C-SVC(Poly=2)	0.776	0.698	0.479
C-SVC(Poly=3)	0.759	0.649	0.395
C-SVC(RBF)	0.612	0.388	0.171
nu-SVC(Linear)	<b>0.798</b>	<b>0.705</b>	<b>0.496</b>
nu-SVC(Poly=2)	0.766	0.671	0.424

nu-SVC(Poly=3)	0.736	0.641	0.384
nu-SVC(RBF)	0.743	0.632	0.373
Decision Tree	0.734	0.519	0.336

#### 4.5 Comparison

We score the concept pairs in dataset 1 and dataset 2 using the proposed semantic similarity measures. Results are shown in Table 4.6 and Table 4.7. Proposed method earns the highest correlation of 0.798 in dataset 1, 0.705 in dataset 2 with physician scores and 0.496 in dataset 2 with expert scores. It shows the highest similarity score for the four concept-pairs including *migraine* and *headache*, *measles* and *rubeola*, *chicken pox* and *varicella*, *down syndrome* and *trisomy 21*. Lowest similarity is reported for *acquired immunodeficiency syndrome* and *congenital heart defects* in dataset 1. It shows the highest similarity score for the four concept-pairs *diabetes mellitus* and *hypertension*. Lowest similarity is reported for *lymphoid hyperplasia* and *laryngeal cancer* in dataset 2.

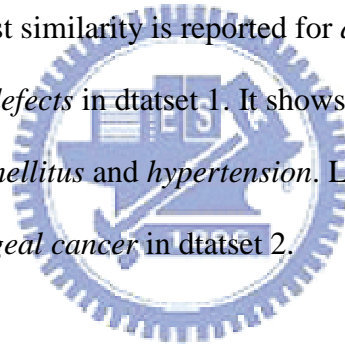


Table 4.6: Dataset 1 with human similarity scores and proposed scores

Concept 1	Concept 2	H	Proposed
Anemia	Appendicitis	0.031	0.697477
Dementia	Atopic Dermatitis	0.062	0.37108
Bacterial Pneumonia	Malaria	0.156	0.444349
Osteoporosis	Patent Ductus Arteriosus	0.156	0.248374
Amino Acid Sequence	Anti-Bacterial Agents	0.156	0.56565
Acquired Immunodeficiency Syndrome	Congenital Heart Defects	0.062	0.210191
Otitis Media	Infantile Colic	0.156	0.520515
Meningitis	Tricuspid Atresia	0.031	0.256254
Sinusitis	Mental Retardation	0.031	0.333204
Hypertension	Kidney Failure	0.5	0.955846
Hyperlipidemia	Hyperkalemia	0.156	0.567689

Hypothyroidism	Hyperthyroidism	0.406	0.999451
Sarcoidosis	Tuberculosis	0.406	0.995609
Vaccines	Immunity	0.593	0.796828
Asthma	Pneumonia	0.375	0.998126
Diabetic Nephropathy	Diabetes Mellitus	0.5	0.950368
Lactose Intolerance	Irritable Bowel Syndrome	0.468	0.883431
Urinary Tract Infection	Pyelonephritis	0.656	0.990715
Neonatal Jaundice	Sepsis	0.187	0.595683
Sickle Cell Anemia	Iron Deficiency Anemia	0.437	0.686173
Psychology	Cognitive Science	0.593	0.999995
Adenovirus	Rotavirus	0.437	0.982612
Migraine	Headache	0.718	1
Myocardial Ischemia	Myocardial Infarction	0.75	0.993638
Hepatitis B	Hepatitis C	0.562	0.999997
Carcinoma	Neoplasm	0.75	0.889407
Pulmonary Valve Stenosis	Aortic Valve Stenosis	0.531	0.960003
Failure To Thrive	Malnutrition	0.625	0.934162
Breast Feeding	Lactation	0.843	0.975854
Antibiotics	Antibacterial Agents	0.937	0.952958
Seizures	Convulsions	0.843	0.999996
Pain	Ache	0.875	0.830473
Malnutrition	Nutritional Deficiency	0.875	0.92306
Measles	Rubeola	0.906	1
Chicken Pox	Varicella	0.968	1
Down Syndrome	Trisomy 21	0.875	1
Correlation		1	0.798

Table 4.7: Dataset 2 with human similarity scores and proposed scores

Concept 1	Concept 2	Phy	Exp	Proposed
Renal Failure	Kidney Failure	4	4	0.975028
Heart	Myocardium	3.3	3	0.910151
Stroke	Infarct	3	2.8	0.924013
Abortion	Miscarriage	3	3.3	0.993801
Delusion	Schizophrenia	3	2.2	0.5
Congestive Heart Failure	Pulmonary Edema	3	1.4	0.998988
Metastasis	Adenocarcinoma	2.7	1.8	0.880069

Calcification	Stenosis	2.7	2	0.747826
Diarrhea	Stomach Cramps	2.3	1.3	0.999967
Mitral Stenosis	Atrial Fibrillation	2.3	1.3	0.962097
Chronic Obstructive Pulmonary Disease	Lung Infiltrates	2.3	1.8	0.349326
Rheumatoid Arthritis	Lupus	2	1.1	0.997619
Brain Tumor	Intracranial Hemorrhage	2	1.3	0.54715
Carpal Tunnel Syndrome	Osteoarthritis	2	1.1	0.8177
Diabetes Mellitus	Hypertension	2	1	0.999998
Acne	Syringe	2	1	0.349637
Antibiotic	Allergy	1.7	1.2	0.849412
Cortisone	Total Knee Replacement	1.7	1	0.279371
Pulmonary Embolus	Myocardial Infarction	1.7	1.2	0.940106
Pulmonary Fibrosis	Lung Cancer	1.7	1.4	0.705904
Cholangiocarcinoma	Colonoscopy	1.3	1	0.351643
Lymphoid Hyperplasia	Laryngeal Cancer	1.3	1	0.241465
Multiple Sclerosis	Psychosis	1	1	0.415343
Appendicitis	Osteoporosis	1	1	0.569876
Rectal Polyp	Aorta	1	1	0.296103
Xerostomia	Alcoholic Cirrhosis	1	1	0.247209
Peptic Ulcer Disease	Myopia	1	1	0.241701
Depression	Cellulitis	1	1	0.375917
Varicose Vein	Entire Knee Meniscus	1	1	NaN
Hyperlipidemia	Metastasis	1	1	0.293352
Correlation				0.705 0.496

Table 4.8 show the results of correlations with human scores for our proposed scores (nu-SVC with 7 features and 900 training samples) using the dataset 1, because we could find only 34 out of the 36 concept pairs in SNOMED-CT as some terms cannot be found, so we used 34 pairs, experimented on SNOMED-CT, and compared with four other measures: SemDist, Path length, Leacock & Chodorow, Wu & Palmer [29], [30], [31],[32]. Our measure achieves the best correlations compared with other four methods.



Table 4.8: Absolute correlations with human scores using SNOMED-CT on dataset 1

SNOMED-CT	
Measure	Dataset 1
SemDist	0.726(2)
Path length	0.422(5)
Leacock & Chodorow	0.6 (3)
Wu & Palmer	0.498(4)
Proposed	<b>0.802(1)</b>

Table 4.9 show the results of correlations with physician and expert scores for our proposed scores (nu-SVC with 7 features and 900 training samples) using the dataset 2, experimented on SNOMED-CT, and compared with six other measures: Path length, Leacock & Chodorow, Lin, Resnik, Jiang & Conrath and Vector(All sect, 1M notes) [30], [31], [33], [34], [35], [36]. Our measure achieves the best correlations with physician scores and fifth best correlations with expert scores compared with other six methods.

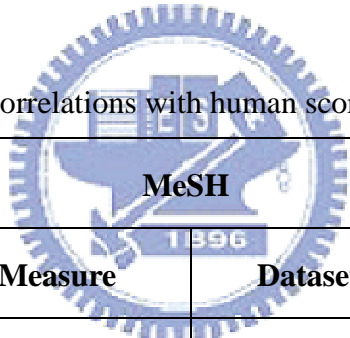
Table 4.9: Absolute correlations with human scores using SNOMED-CT on dataset 2

SNOMED-CT		
Measure	Dataset 2(Phy)	Dataset 2(Exp)
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)	0.61(3)

Jiang & Conrath	0.506(5)	<b>0.741(1)</b>
Vector(All sect, 1M notes)	0.436(6)	0.497(5)
Proposed	<b>0.706(1)</b>	0.496(6)

Table 4.10 show the results of correlations with human scores for our proposed scores ( nu-SVC with 7 features and 900 training samples) using the dataset 1, experimented on MeSH, and compared with eleven other measures: SemDist, Path length, Leacock & Chodorow, Wu & Palmer, Lin, Jiang & Conrath, Resnik, Li, Lord, Tversky, Rodriguez [29], [30], [31] , [32], [33] , [35], [34] , [37, [38] , [15], [39]. Our measure achieves the fourth best correlations compared with other eleven methods.

Table 4.10: Absolute correlations with human scores using MeSH on dataset 1



<b>MeSH</b>	
<b>Measure</b>	<b>Dataset 1</b>
SemDist	<b>0.825(1)</b>
Path length	0.765(5)
Leacock & Chodorow	0.82(2)
Wu & Palmer	0.811(3)
Lin	0.723(6)
Jiang & Conrath	0.71(8)
Resnik	0.718(7)

Li	0.705(9)
Lord	0.701(10)
Tversky	0.67(11)
Rodriguez	0.69(12)
Proposed	0.798(4)



Table 4.11 show the results of correlations with physician and expert scores for our proposed scores ( nu-SVC with 7 features and 900 training samples) using the dataset 2, because we could find only 25 out of the 30 concept pairs in SNOMED-CT as some terms cannot be found, so we used 25 pairs experimented on MeSH, and compared with five other measures: SemDist, Path length, Leacock & Chodorow, Wu & Palmer, Choi & Kim [29], [30], [31] , [32], [40]. Our measure achieves the best correlations with physician scores and sixth best correlations with expert scores compared with other five methods.

Table 4.11: Absolute correlations with human scores using MeSH on dataset 2

<b>MeSH</b>		
<b>Measure</b>	<b>Dataset 2(Phy)</b>	<b>Dataset 2(Exp)</b>
SemDist	0.666(3)	<b>0.863(1)</b>
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	0.56(6)	0.724(5)
Proposed	<b>0.723(1)</b>	0.539(6)

## Chapter 5 Conclusions and Future Work

### 5.1 Conclusions

In this paper, we proposed a measure that utilizes page counts to calculate semantic similarity robustly between two given concepts or terms. The method consists of fifteen features apply support vector machines and decision tree classifier models. Training data were manually collected from two websites: MedicineNet.com & synonyms.net. Proposed method outperformed all the baselines on two datasets. A high correlation coefficient 0.798 with human ratings was found for semantic similarity on the dataset provided by A. Hliaoutakis. With physician's ratings, correlation coefficient of 0.705 was found for semantic similarity on the dataset provide by T. Pedersen et al; the correlation coefficient of 0.496 with expert's ratings was found. Only 7 features and 900 training samples are necessary to leverage the proposed method using nu-SVC with linear kernel. A contrasting feature of our method compared to the ontology- based semantic similarity measures is that our method requires no taxonomies, such as SNOMED-CT or MeSH, for calculation of similarity. Therefore, the proposed method can be applied in many tasks where taxonomies are not up-to-date or do not exist. We also realize that our study measures produce much closer correlations with physician scores than those with medical experts. However, all the ontology measures are reversed.

### 5.2 Future Work

Further study can be summarized:

- We can enhance the models by using more lexico-syntactic patterns that can capture the synonymous concept pairs more precisely.
- We can use another feature selection strategy to increase the accuracy.
- We intend to apply the proposed semantic similarity measure in automatic synonym extraction, query suggestion and name alias recognition.

## References

- [1] S. Alexaki, V. Christophides, G. Karvounarakis, D. Plexousakis, K. Tolle, B. Amann, I. Fundulaki, M. Scholl, and A.-M. Vercoustre. Managing RDF Metadata for Community Webs. In *Proceedings of the ER'00 2nd International Workshop on the World Wide Web and Conceptual Modeling (WCM'00)*, pages 140{151, Salt Lake City, Utah, 9-12 October 2000.
- [2] Grigoris Antoniou and Frank van Harmelen. *A Semantic Web primer*. 2004.
- [3] D.L. McGuinness, R. Fikes, J. Rice, and S. Wilder. An Environment for Merging and Testing Large Ontologies. In *Proceedings of the 7th International Conference on Principles of Knowledge Representation and Reasoning (KR'00)*, Breckenridge, Colorado, USA, 12-15 April 2000.
- [4] D.L. McGuinness, R. Fikes, J. Rice, and S. Wilder. The Chimaera Ontology Environment. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI'00)*, Austin, Texas, 30 July - 3 August 2000.
- [5] S.J. Nelson, D. Johnston, and B.L. Humphreys. Relationships in Medical Subject Headings. In C.A. Bean and R. Green, editors, *Relationships in the Organization of Knowledge*, pages 171{184. Kluwer Academic Publishers, New York, 2001.
- [6] W. Douglas Johnston Stuart J. Nelson and Betsy L. Humphreys. Relationships in Medical Subject Headings (MeSH). In *National Library of Medicine, Bethesda, MD, USA*, 2002.
- [8] P.R. Cohen and R. Kjeldsen. Information Retrieval by Constrained Spreading Activation in Semantic Networks. *Information Processing and Management*, 23(4):255{268, 1987.
- [9] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17{30, January/February 1989.
- [11] Z. Wu and M. Palmer. Verb Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics (ACL'94)*, pages 133{138, Las Cruces, New Mexico, 1994.
- [12] Yuhua Li, Zuhair A. Bandar, and David McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871{882, July/August 2003.
- [13] O. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. *Journal of Artificial Intelligence Research*, 11:95{130, 1999.
- [14] J.J. Jiang and D.W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistic*, Taiwan, 1998.
- [15] A. Tversky. Features of Similarity. *Psychological Review*, 84(4):327{352, 1977.
- [16] M.A. Rodriguez and M.J. Egenhofer. Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*,

15(2):442-456, March/April 2003.

[17] Bollegala, D., Matsuo, Y., Ishizuka, M. Measuring Semantic Similarity between Words using Web Search Engines. In: Proc. Int. WWW2007 Conf., 2007.

[18] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. In Proceedings of 15<sup>th</sup> International World Wide Web Conference, 2006.

[19] Rudi L. Cilibrasi and Paul M.B. Vitányi. The Google Similarity Distance IEEE ITSOC Information Theory Workshop 2005 on Coding and Complexity, 29<sup>th</sup> Aug. - 1st Sept., 2005,

[20] John McCrae\* and Nigel Collier. Synonym set extraction from the biomedical literature by lexical pattern discovery *BMC Bioinformatics* 2008, 9:159 doi:10.1186/1471-2105-9-159

[21] Yi-Wei Chen and Chih-Jen Lin. Combining SVMs with Various Feature Selection Strategies

[22] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61-74, 2000.

[23] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

[24] Chih-Chung Chang and Chih-Jen Lin, Training  $\nu$ -Support Vector Classifiers: Theory and Algorithms.

[25] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

[26] K. M. Sim and P. T. Wong, "Toward agency and ontology for web-based information retrieval," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 257-269, Aug. 2004.

[27] A. Hliaoutakis, "Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline," Master's thesis, Tech. Univ. Crete, Chaniá, Crete, 2005.

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

[29] Hisham Al-Mubaid, and Hoa A. Nguyen. Measuring Semantic Similarity Between Biomedical Concepts Within Multiple Ontologies IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 39, NO. 4, JULY 2009.

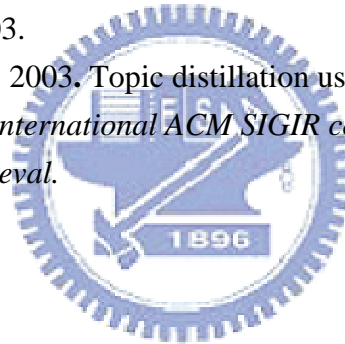
[30] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 17-30, Jan./Feb. 1989.

[31] 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: MIT Press, 1998, pp. 265-283.

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

[33] Lin D. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Madison, WI; 1998. p. 296-304.

- [34] Resnik P. WordNet and class-based probabilities. In: Fellbaum C, editor. WordNet: An electronic lexical database. Cambridge, MA: MIT Press; 1998. p. 239–63.
- [35] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th international conference on research in computational linguistics, Taipei, Taiwan; 1997. p. 19–33.
- [36] Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan, Christopher G. Chute. Measures of semantic similarity and relatedness in the biomedical domain *Journal of Biomedical Informatics* 40 (2007) 288–299.
- [37] Yuhua Li, Zuhair A. Bandar, and David McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871{882, July/August 2003.
- [38] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation. *Bioinformatics*, 19(10):1275{83, 2003.
- [39] M.A. Rodriguez and M.J. Egenhofer. Determining Semantic Similarity Among Entity Classes from Di@erent Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442{456, March/April 2003.
- [40] Ikkyu Choi & Minkoo Kim. 2003. Topic distillation using hierarchy concept tree. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*.





## Appendix

### A. Google AJAX Search API

The following command performs a Web search (/ajax/services/search/web), for Kidney Failure (q=Kidney%20Failure).

```
curl -e http://www.my-ajax-site.com \  
'http://ajax.googleapis.com/ajax/services/search/web?v=1.0&q= Kidney% 20Failure '
```

The response has a Content-Type of text/javascript; charset=utf-8. The response below that the responseData is identical to the results.

```
{ "responseData": {  
  "results": [  
    {  
      "GsearchResultClass": "GwebSearch",  
      "unescapedUrl": "http://en.wikipedia.org/wiki/Kidney_Failure",  
      "url": "http://en.wikipedia.org/wiki/Kidney_Failure",  
      "visibleUrl": "en.wikipedia.org",  
      "cacheUrl":  
      "http://www.google.com/search?q\u003dcache:TwrPfh22hYJ:en.wikipedia.org",  
      "title": "\u003cb\u003eKidney Failure\u003c/b\u003e - Wikipedia, the free  
      encyclopedia",  
      "titleNoFormatting": "Kidney Failure - Wikipedia, the free encyclopedia",  
      "content": "[1] In 2006, she released her debut album..."  
    },  
    {  
      "GsearchResultClass": "GwebSearch",  
      "unescapedUrl": "http://www.imdb.com/name/nm0385296/",  
      "url": "http://www.imdb.com/name/nm0385296/",  
      "visibleUrl": "www.imdb.com",  
      "cacheUrl":  
      "http://www.google.com/search?q\u003dcache:1i34Kkqns00J:www.imdb.com",
```

```

"title": "\u003cb\u003eKidney Failure\u003c/b\u003e",
"titleNoFormatting": "Kidney Failure",
"content": "Self: Zoolander. Socialite \u003cb\u003eKidney Failure\u003c/b\u003e..."
},
...
],
"cursor": {
"pages": [
{ "start": "0", "label": 1 },
{ "start": "4", "label": 2 },
{ "start": "8", "label": 3 },
{ "start": "12", "label": 4 }
],
"estimatedResultCount": "286000",
"currentPageIndex": 0,
"moreResultsUrl":
http://www.google.com/search?oe\u003dutf8\u0026ie\u003dutf8...
}
}
, "responseDetails": null, "responseStatus": 200}

```



**B. 100 synonymous concept pairs of training data**

Concept 1	Concept 2
Abdomen	Tummy
Abducens Nerve	Abducent Nerve
Abortifacient	Aborticide
Achondroplasia	Achondroplasty
Achromycin	Tetracycline
Bacteria	Bacterium
Basophil	Basophile
Bedsore	Decubitus Ulcer
Benzene	Benzol
Blastoma	Embryonal Carcinosarcoma

Calamine	Hemimorphite
Carbohydrate	Saccharide
Cardiac murmur	Heart murmur
Catabolism	Destructive Metabolism
Cardiovascular system	Circulatory system
Breakbone Fever	Dandy Fever
Dermis	Derma
Diaphragmatic Hernia	Hiatal Hernia
Dizziness	Giddiness
Dropsy	Eedema
Echinococcosis	Hydatidosis
Ectopic pregnancy	Ectopic gestation
Electrocardiogram	Cardiogram
Electrophoresis	Dielectrolysis
Facial Nerve	Seventh Cranial Nerve
Farsightedness	Longsightedness
Fascioliasis	Fasciolosis
First Cranial Nerve	Olfactory Nerve
Fistula	Fistulous Withers
Gallus Gallus	Red Jungle Fowl
Gamma Radiation	Gamma Ray
Gangrene	Necrosis
Gargoylism	Lipochoondrodystrophy
Genital Wart	Venereal Wart
Goiter	Struma
Hallucination	Delusion
Heat Prostration	Heat Exhaustion
Hemochromatosis	Iron Overload
Hepatocarcinoma	Hepatocellular Carcinoma
Herpes Genitalis	Genital Herpes
Heterosexuality	Heterosexualism
Ileus	Intestinal Obstruction
Implantation	Nidation
Infant	Babe
Inguinal Canal	Canalis Inguinalis
Intersex	Androgyne
Intestine	Gut

Iodine	Iodin
Iontophoresis	Electromotive Drug Administration
Joint	Articulatio
Jaundice	Icterus
Keloid	Cheloid
Kinetosis	Motion sickness
Knee	Genu
Kyphosis	Humpback
Labyrinthitis	Otitis Interna
Lachrymal Gland	Lacrimal Gland
Lactase Deficiency	Lactose Intolerance
Lateral Epicondylitis	Tennis Elbow
Leishmaniasis	Leishmaniosis
Lienal Artery	Splenic Artery
Limb	Arm
Lymphopathia Venereum	Lymphogranuloma Venereum
Male Erectile Dysfunction	Erectile Dysfunction
Malignant Hepatoma	Hepatocellular Carcinoma
Mandibular Joint	Temporomandibular Joint
Mediterranean Anemia	Thalassaemia
Medication	Medicament
Meiosis	Miosis
Melasma	Chloasma
Nausea	Sickness
Necrobiosis Lipoidica Diabeticorum	Necrobiosis Lipoidica
Neocortex	Neopallium
Nephrolith	Kidney Stone
Nervus Glossopharyngeus	Glossopharyngeal Nerve
Neurogliaocyte	Glial Cell
Oesophageal Reflux	Esophageal Reflux
Onchocerciasis	River Blindness
Orthodontics	Orthodontia
Paleostriatum	Pallidum
Palpebra	Eyelid
Parasite	Sponge
Pars Nervosa	Posterior Pituitary
Pedigree	Ancestry

Periodontitis	Periodontal Disease
Peristalsis	Vermiculation
Phenol	Phenylic Acid
Pituitary Gland	Pituitary Body
Plastic Surgery	Reconstructive Surgery
Rachischisis	Spina Bifida
Rale	Rattle
Regional Ileitis	Regional Enteritis
Restriction Enzyme	Restriction Endonuclease
Retinal Detachment	Detachment of the Retina
Sandfly Fever	Pappataci Fever
Seborrheic Dermatitis	Seborrheic Eczema
Second Cranial Nerve	Optic Nerve
Shingles	Zoster
Sixth Cranial Nerve	Abducent Nerve
Tenth Cranial Nerve	Wandering Nerve
Third Cranial Nerve	Oculomotor Nerve
Tympanic Cavity	Middle Ear
Uterus	Womb

**C. 100 non-synonymous concept pairs of training data**

Concept 1	Concept 2
Abdomen	Awareness
Abortifacient	Cramp
Absinthe	Absinthe
Acathisia	Odor
Aflatoxin	Smallpox
Adenovirus	Somnambulism
Adventitia	Spine
Bacteriophage	Humpback
Balantidiasis	Keloid
Balantidium	Keratin
Bariatrics	Kindred
Barotrauma	Kinship
Bedsore	Kyphosis
Beriberi	Lactation
Bevacizumab	Lassitude

Cadaver	Caudate Nucleus
Caffeine	Cavernous Hemangioma
Calamine	Cavernous Sinus
Cavernous Sinus Thrombosis	Calcaneus
Celiac Disease	Herpes Zoster
Developmental Delay	Impact
Developmental Disorder	Implantation
Diabetic Ketoacidosis	Impotence
Diabetic Nephropathy	Incision
Diabetic Neuropathy	Incubator
Ear Piercing	Attention Deficit Disorder
Ear Wax	Attention Deficit Hyperactivity Disorder
Eastern Equine Encephalitis	Hypoglossal Nerve
Ebola Virus	Pituitary Gland
Ectodermal Dysplasia	Intestinal Obstruction
Facies	Suckling
Family	Sunspot
Farsightedness	Travel
Fascia	Tug
Fasciculation	Vertigo
Gait	Inflammation
Gait	Coated Stent
Galactorrhea	Inflammation
Galactose	Inflammation
Galactosemia	Injury
Habitual Abortion	Dermis
Hair Follicle	Development
Hallucination	Fontanel
Hallucination	Cyclic Citrullinated Peptide
Hallucinogen	Fontanelle
Ibuprofen	Cholesterol
Ichthyosis	Cilium
Ichthyosis Vulgaris	Middle Ear
Icterus	Clavicle
Idiopathic Pulmonary Fibrosis	Middle Ear
Jaundice	Deglutition

Jaundice	Sandhoff Disease
Jaw	Deglutition
Jejunostomy	Dehydration
Jejunum	Ergocalciferol
Kaposi Sarcoma	Laparoscopic Cholecystectomy
Kartagener Syndrome	Tennis Elbow
Karyokinesis	Impairment
Karyotype	Innovation
Keloid	Knowledge
Lab	Ataxy
Labia	Atmosphere
Labor	Atrophy
Labor	Scleredema Adultorum
Labyrinth	Audiometry
Macrobiotic Diet	Tick Fever
Macular Hole	Oral Cavity
Meibomian Gland	Seventh Cranial Nerve
Magnesium Deficiency	Kawasaki Disease
Magnetic Resonance Elastography	Cystic Fibrosis
Naegleria Fowleri	Acoustic Nerve
Nasal Septum	Eighth Cranial Nerve
Natriuretic Peptide	Glial Cell
Natural Immunity	Glial Cell
Natural killer cell	Posterior Pituitary
Obstetrical Forceps	Renal pelvis
Occipital Bone	Hip Joint
Occupational Medicine	Proteolytic Enzyme
Oculocutaneous Albinism	Periodontal Disease
Olfaction	Tarsal Tunnel Syndrome
Paleostriatum	Tetralogy of Fallot
Parasite	Thanatophoric dysplasia
Parasitemia	Therapeutic Touch
Paresthesia	Thoracic Aorta
Paroxysm	Thoracic Duct
Quackery	Thyroid Cartilage
Quiescence	Thyroid Hormone Receptor
Rabies	Accessory

Race	Accessory
Radial Keratotomy	Prostate Cancer
Radiation	Accessory
Radiation	Thyroid Stimulating Hormone
Radiation Fibrosis	Roseola Infantum
Radical Neck Dissection	Temporal Lobe Epilepsy
Sabin Vaccine	Sebaceous Gland
Saccular Aneurysm	Seborrheic Dermatitis
Safe Sex	Optic Nerve
Salivary Gland	Testicular Cancer
Salk Vaccine	Blood Poisoning
Xerostomia	Lung Cancer
Yerba Mate	Knee Joint

