AUTOMATIC HIERARCHICAL CLASSIFICATION OF DOCUMENTS

A dissertation submitted in partial fulfillment of

the requirements for the degree

of

Master of Technology
in
Computer Technology

by

M.Kalaiselvi Geetha

(2000EET014)

Under the guidance of

Dr. Lipika Dey

Department of Electrical Engineering
Indian Institute of Technology, Delhi, India
December 2001

CERTIFICATE

This is to certify that the dissertation titled **"AUTOMATIC TAXONOMICAL CLASSIFICATION OF DOCUMENTS "** is a bona fide work done by **Mrs. M.KalaiSelvi Geetha** for the partial fulfillment of the requirement for ***Major project*** in the ***Department of Electrical Engineering*** at ***Indian Institute of Technology, Delhi***. This project work was carried out under my guidance and has not been submitted elsewhere.

Prof. Lipikadey,
Mathematics Dept.
IIT, Delhi

Dedicated To....

.... My Beloved Children....

Abhi & Khiran

ACKNOWLEDGEMENTS

After paying a humble obeisance to the benign almighty, I would like to express my deep gratitude to my advisor, **Dr. Lipika dey** for her wholehearted involvement, co-operation, guidance and patience which made possible the realization of my efforts. Her guidance, encouragement, highly positive attitude towards the project, frequent discussions for the project have made this project a success.

I express my sincere thanks to **Prof. Santanu choudhury** for permitting me to carry out my project work in Computer Technology Lab.

I convey my thanks to **Mr. Fateh Singh** of the Computer Technology Lab who has always been kind hearted and ever ready to help me in carrying out work in the Lab.

I am indebted to my parents **Mr. N. Manoharan** and **Mrs. Gunajothy Manoharan** for the confidence they have developed in me since childhood to work hard. This work would not have been completed without the encouragement of my husband**, Mr. B. Sivaraman** and my brother **Mr. M. Murali** who gave me moral support during the hard time I faced in the M.Tech degree. I would like to convey my gratefulness to my in-laws **Mr. Balasubramanian** and **Mrs. Suseela Balasubramanian** who supported me to pursue M.Tech by taking care of my children.

I dedicate this thesis to my children, **Baby. C. S. Abhinaya** and **Master. C.S. Khiran Kumar**, who readily understood and adjusted my absence during my stay away from home. Their love will always inspire me to achieve new goals.

CONTENTS

ABSTRACT

Text classification is becoming increasingly more important with the proliferation of the Internet and the large amount of data it transfers. Internet Directories, Digital Libraries, Patent Databases are manually organized into topic hierarchies called taxonomy that are maintained manually. However, the exponential growth in the volume of online data makes it impractical to maintain such hierarchies manually. Accordingly, the need for automatic and precise classifier arises. We have proposed a method for constructing automatic hierarchical classifiers that uses Fisher's Discriminant to separate the feature words at each level in the taxonomy. Using the feature words, the document signatures are extracted out. When new documents are added to the repository, their position in the hierarchy is estimated using Bayesian conditional probability. The system works with free form text and has been shown to have accuracy of around 80% or more in a number of domains.

CHAPTER 1

Introduction

1.1 Preface:

The amount of online data in the form of free format text is emerging tremendously swiftly. Information retrieval from this vast monolithic collection is an overwhelming task. Consequently, there is an escalating need for proper organization of the available data in an efficient structure that enhances effective and pertinent information retrieval. The first step towards organization is classification of the data by distinguishing them and to put the relevant data under an appropriate category. This classification structure is also called as Taxonomy. One can then excavate the desired data that was classified under each node in the Taxonomy.

1.2 Problem Definition:

The text databases on the web are 'hidden' behind search interfaces, and their documents are only accessible through querying. Search engines in general exhibit the contents of such search-only databases. They work mainly with the presence or absence of keywords in the query while searching these databases. Recently, Yahoo! like directories have started to manually systematize these databases into categories that users can browse to find these valuable resources. However, the categorization is done manually and organizing millions of databases available in the Web by hand is not a simple job. At this juncture, we propose a method for automatic and hierarchical classification of data that enhances the effectiveness and relevance of information retrieved.

1.3 Requirements Of The Classification System:

Classification systems have long been used to give structure to large bodies of information. A well-formulated system can improve understanding of the information as well as provide easy access to it, thus making the information more useful. As text repositories grow in number and size and global connectivity improves, there is a pressing need to support efficient and effective information retrieval (IR), search and filtering. It is common to manage complexity by hierarchy. To facilitate information retrieval in text using hierarchy, we have to build a system that enables search and navigation in taxonomies aided by proper classification of documents. In order to build a system that enables search and navigation in taxonomies, the subsequent requirements are to be met.

1. Apart from keywords, documents loaded into databases must be indexed on topic paths in the taxonomy. A reliable automatic hierarchical classifier is needed to achieve this. As one goes deep into the taxonomy, shared jargon makes automatic topic separation difficult.
2. The taxonomy should be also used to present to the user a series of progressively refined views of document collections in response to user's queries.
3. The system must be accurate.
4. The system must efficiently update its knowledge. After significant number of new documents has been added to the taxonomy, same classification criterion may not remain valid. Also, appropriate corrections have to be introduced, when it makes mistakes and a human intervenes.

1.4 Aim Of The Project:

The aim of the project was to develop a method for automatic hierarchical classification of HTML documents into a defined taxonomy. The classifier was built after analyzing a pre-defined hierarchy with a set of training documents. We have used a Fisher's discriminant based scheme to form a hierarchical classifier. We have worked with multiple domains and established the correctness of the approach with results obtained over a large test set in each of the domains. We have also worked on efficient retrieval of text information, given a search query. The uniqueness of the

approach lies in the fact that while some classifiers use only a limited number of words picked up from abstracts, keywords and index-words, our classifier works with different types of dominant structures extracted from free form HTML documents in general.

1.5 Layout Of The Thesis:

Chapter 2 explains the general overview of search engines. This chapter explains how Internet search engines search for relevant documents. Chapter 3 starts with brief understanding of the hierarchical structure. We have then presented the proposed classification system. It also talks about the context sensitive signatures and context sensitive feature selection. It then focuses on automatic classification of documents. Chapter 4 gives a detailed overview of the system implementation, and the results obtained are presented.

CHAPTER 2

General Overview Of Search Engines

2.1 Introduction:

The Internet and its most discernible component, the World Wide Web, store hundreds of millions of pages of information, waiting to present information on an amazing variety of topics. However most of these pages are titled according to the whim of their authors, and are stored on servers with cryptic names. Search engines accept queries from users and return addresses of sites where relevant documents are stored.

2.2 How Do The Internet Search Engine Works:

Internet search engines are unique sites on the Web that are designed to rally round people to find information stored on other sites. There are differences in the ways various search engines work, but they all perform three basic tasks:

- They search the Internet or select pieces of the Internet based on important words.
- They keep an index of the words they find, and where they find them.
- They allow users to look for words or combinations of words found in that index.

2.3 Looking At The Web:

Before a search engine can state where a file or document is, it must be found. To find information on the hundreds of millions of Web pages that

exist, a search engine employs special software robots called spiders, to build lists of the words found on Web sites called Web Crawling. The usual starting points of the spiders are lists of very popular pages. The spider will begin with a popular site, indexing the words on its pages and following every link found within the site, thus spreading out across the most widly used portions of the web.

2.4 Building An Index:

Once the spiders have completed the task of finding information on Web, indexing is performed. In this stage, each document is rehabilitated into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size and capitalization. In addition, all the links in every web page are parsed out and imperative information about them is stored in anchors file.

2.5 Building A Search:

Searching through an index involves a user building a query and submitting it through the search engine. The query can be quite simple, a single word at minimum to complex query with the use of Boolean operators that allow you to refine and extend the terms of the search.

2.6 Current Search Engines:

Before discussing about the search engines, we define the terms, recall and precision.

2.6.1 Recall:

The recall of documents A with respect to the total reference Ar is defined as:

$$\text{Recall} = |A \text{ inter } Ar| \; / \; |Ar|$$

Eqn (2.1)

'inter' is the set-theoretical intersection.

It represents the proportion of documents A that are correct with respect to the total reference Ar.

2.6.2 Precision:

The precision of documents A with respect to the total reference Ar is defined as:

$$\text{Precision} = |A \text{ inter } Ar| / |A|$$

Eqn (2.2)

It represents the proportion of documents A that are right with respect to the total of those proposed.

2.6.3 Search Engines:

Many search engines like Yahoo!, AltaVista, Google etc., are crawling on the web. Most queries posted to search engines are very short. Such queries routinely suffer from abundance problem. In Yahoo!, the documents are classified according to keywords, but not all relevant terms are covered. In Google or AltaVista, every single occurrence of a term in a document is indexed, no matter how badly it represents the document content. The coverage is higher since all terms are considered but this slows down the system.

AltaVista works with a flat unstructured set of classes. The recall is high but precision is low as irrelevant material is also indexed. Yahoo! works on man made taxonomy where the precision is high but the manual classification is a daunting problem. Because of its precision, Yahoo! is more extensively used nowadays.

2.7 Future Search:

The search engines look for the words or phrases exactly as they are entered as a flat structure. This can be a problem when the entered words have multiple meanings. For example, “Bed” can be a place to sleep, a place where flowers are planted, the storage space of a truck or a place where fish lay their eggs. If “Bed” is fed as a query, pages featuring all of the others are also listed.

The performance of search engines can be improved by organization of the available information in a hierarchical structure that facilitates the reference to concepts and relationships. Users can retrieve more relevant information from hierarchical categories rather than keywords.

2.8 Basic Hierarchical Structure:

A hierarchy describes how data is organized within a system. The basic structure of a classification of plants and animals hierarchy is shown in the figure (1). Taking plants and animals classification as a reference; the hierarchical structure clearly brings out their classification. If the documents matched against the toplevel node, i.e Animal and Plants Classification is found to have a significant measure of similarity it would subsequently be matched against the two subclasses, Plants and Animals. They are in turn classified further to their subclasses, if similarity is found. The plants are classified under two categories (i) Structure and (ii) Reproduction where, Structure includes documents that describes how the plants suck water and Reproduction includes documents that explains the reproductive method of plants. Further, under Structure the documents are classified as Vascular and non-Vascular. In Reproduction, the documents are classified as Seeds and Spores. Under animals, they are classified into two categories, (i) Vertebrates and (ii) Invertebrates where, Vertebrates includes documents that explain about the animals with backbones. Whereas, Invertebrates includes documents that gives details of animals without backbones. They are further classified as Carnivores, Herbivores and Omnivores in case of Vertebrates, and Soft bodied, Mollusks, Exoskeletons in case of Invertebrates. The documents are placed at the lowermost level and the upper nodes correspond to the various categories and subcategories in which documents have been classified.

2.9 Summary:

Current search engines crawl the web as a flat unstructured data. Hence much irrelevant information is retrieved. For that reason, we need a structured classification of data. At this point, a method based on hierarchical classification as in library is proposed and is presented in chapter 3.

CHAPTER 3

Searching Within Taxonomy

3.1 Introduction:

Taxonomies are structures that provide a way of classifying things into a series of hierarchical groups to make them easier to identify, study or locate. Taxonomic classification focuses on the purpose of user's search, the method user use to find the information and the content of the information of the search. For building the automatic hierarchical classifier, we aim to get context sensitive feature terms for each node of the taxonomy on the basis of which, we proceed further in the classification to build the multilevel classifier.

In order to build an automatic hierarchical classifier, we first assume that when a system is given a set of documents and their hierarchical classification, then the system can learn from these examples, the principle of classification. As well as a measure for importance of words in the context of a class has to be somehow determined, since all automatic classifications have to be done only on the basis of words. We use techniques from pattern recognition to separate the feature words or discriminants from the noise words efficiently at each node of the taxonomy. Furthermore, if one wants to go for progressive classification, subsequently it is also required to learn the discriminating power of words at a class. Hence, in our context learning the system implies finding the most discriminating terms for each node for a given set of preclassified documents. When classifying new documents, only the feature terms are used. Good features are few in number, so the class models are small and the classifications are accurate. Once the system learns the principle of classification, it will be possible for it to apply the acquired knowledge to classify new documents appropriately. In this chapter, we will present in detail the techniques that make possible the capabilities mentioned above.

3.2 Multilevel Classifier:

3.2.1 Steps:

To build a multilevel classifier we did the following:

- Downloaded a set of documents and classified them by hand.
- Extracted important terms from these.
- Extracted signatures for classes as collections of words with high Fisher's discriminant value.

Later, we used the classifier for automatic classification of documents according to the hierarchy.

3.2.2 Term Extraction:

For extracting terms from the downloaded documents, importance must be given to the nature of terms, like title words, keywords, capitalization, font size etc., we set up the following strategy.

$$x(d,t) = \frac{\sum_i w_i \, n_i(d,t)}{n(d)}$$

Eqn (3.1)

Where, $x(d,t)$ is the term to be extracted.

w_i is the weight associated to the nature of terms.

$\frac{n_i(d,t)}{n(d)}$, is the occurrence rate of t in d.

$n(d,t)$ is the occurrence of t in d.

$n(d)$ is the number of terms in d.

Once, the terms are extracted out, the importance of the terms with respect to the corpus has to be determined. For that, we should decide, whether the term is a feature term or a noise term in that corpus.

3.2.3 Feature And Noise Terms:

When building a model for each class c from a training set, we must decide if a term t appears only incidentally, or sufficiently consistently to suspect a causal connection; t is accordingly a *noise term* or a *feature term*. Given a new document, we should focus our attention only on the features for classifying it.

We are constrained in both ways: we cannot miss the highly discriminating terms, and we cannot include everything, because the frequencies of some terms are noisy and unindicative of content. This is called *feature selection* problem. Roughly speaking, we are in search of a set of terms that minimizes the probability that a document is misclassified. We therefore restrict ourselves to the following approach: first we assign a merit measure to each term, then pick a prefix of terms with highest merit to select the measure of the term, we use *Fisher's Discriminant*. We define some of the related terms and the *Fisher's Discriminant* Function used in the subsequent discussion.

3.2.3.1 Discriminant Analysis:

Discriminant analysis is commonly used to determine which variables discriminate best between two or more groups. The basic idea underlying discriminant analysis is to determine whether groups differ with regard to the mean of a feature variable. This variable is then used to predict group membership.

Discriminant analysis is a very useful statistical tool. It takes into account the different variables of an object and works out which group the object most likely belongs to. Discriminant Analysis is a very useful tool for:

1. Detecting the variables that allow the user to discriminate between different naturally occurring groups.
2. Classifying cases into different groups with a better than chance accuracy.

3.2.3.2 Principal Component Analysis:

Principal component analysis is a useful tool for categorization of data, since it separates the dominating features in the data set. The objectives of Principal Component Analysis is,

- To discover or to reduce the dimensionality of the data set.
- To identify new meaningful underlying variables.

3.2.3.3 Fisher's Discriminant:

To build the automatic hierarchical classifier, we need to search for the best feature set. However, what the best possible classifier does is not known, since there are too many terms in the lexicon. Hence, in practice, we are interested in doing this for our fixed classifier. We want a heuristic that is essentially linear in the original number of terms. Consequently, we first assign a merit measure to each term, and then pick a prefix of terms with highest merit. The merit measure is assigned based on *Fisher's Discriminant.*

Fisher's discriminant is a classification method that projects high-dimensional data onto a line and performs classification in this one-dimensional space. Suppose, we are given two sets of points in n-dimensional Euclidean space, interpreted as two classes, Fisher's discriminant finds a direction on which to project all the points so as to maximize (in the resulting one-dimensional space) the relative class separation as measured by the ratio of inter-class to intra-class variance.

Now, we will discuss how to select the measure for the terms that the classifier will use in its models. Suppose we are given two sets of points in n dimensional Euclidean space, interpreted as two classes. Fisher's method finds a direction on which to project all the points so as to maximize the resulting class separation as measured by the ratio of inter class to intra class variance. In general, given a set of two or more classes {c}, with | c | documents in class c, we compute the ratio of the between class to within class scatter. We express this as,

$$\text{Fisher}(t) = \frac{\sum_{c1,c2} \left[\mu(c_1,t) - \mu(c_2,t)\right]^2}{\sum_{c} (1/|c|) \left[x(d,t) - \mu(c,t)\right]^2}$$

Eqn (3.2)

Where, x(d,t) is defined as in eqn(3.1)

$$\mu(c,t) = (1/|c|) \sum_{d \in c} x(d,t)$$

c_1, c_2 are the classes in the system.

$|c|$ is the documents in class c

The numerator in eqn(3.2) gives the between class scatter and the denominator gives the within class scatter.

3.3 Querying In Taxonomy:

In taxonomy, a useful signature is a function of both the document and reference node. Therefore, significant improvement in search quality may be obtained by maintaining functionally separate indices at each taxonomy node, using only a few signature terms from each document. As a result, now we are having documents that associate not only keywords, but also document topics. The query can now be given depending on the categories instead of responding with the complete flat structured dataset. Hence the user can pose queries by concept, not by keywords. This is done with the help of the following two concepts.

3.3.1 Context Sensitive Signatures:

AltaVista's exhaustive keyword index is more of a problem than a solution. A single occurrence of a term in a document, no matter how useless an indicator of the contents is indexed. So, there is a need to find prototypes that extract signature terms, which are then used for indexing. These signatures can also be used as summaries or thumbnails. Their descriptive power can often compare favorably with that of arbitrary sentences as extracted by popular search engines. In case of taxonomy, a useful signature is a function of both the document and the reference node. The signature includes terms that are 'surprising' given the path from the root to the reference node.

3.3.2 Context Sensitive Feature Selection:

Separating feature terms from noise terms is central to all of the capabilities we have talked about. Now, this is done in following two steps.

- We filter the stop words like was, is ,above it etc.,
- We try to find relevant terms at each node, which are capable of representing documents belonging to certain child class and we call them feature terms of that class.

Stop words have high frequency but low utility. Corpus utilized to eliminate stop words, but it is tricky to handcraft the stop words out of domain knowledge of the language. For example, 'can' is frequently included in stop word lists, in a corpus on waste management, it is a feature term. Thus contents of a stop word list should be highly dependent on the corpus. In hierarchical categories, the importance of a search term depends on the position in the hierarchy. Feature selection enables fine-grained classification on taxonomy. For diverse top-level topics, a single step classifier suffices. But as a document is routed deep into a taxonomy, shared jargon makes sophisticated feature selection important. Together with feature selection, we have to pick models for each class and a classifier. The technique used for this purpose is based on calculating the Fisher's discriminant for each term at each node using the eqn(3.2). The more discriminating terms get high fisher values.

3.4 Selecting A Cut Off:

Let F be the list of terms in our lexicon sorted by the decreasing Fisher index. Our heuristic is to pick from F a prefix F_k of the k most discriminating terms. F_k must include most features and exclude most noise terms. A short F_k enables fast classification and holding a larger taxonomy in memory. Too large an F_k will fit the training data very well but will result in degraded accuracy for test data due to over fitting. Therefore, we sort F_k with decreasing Fisher index and we select only top 80 to 100 values at every node. The number of Fisher index chosen may vary depending on the number of documents at each node.

3.5 Document Model for Classification:

Existing classifiers deal with flat set of classes. Whereas, in our model, the feature set changes by context as the document proceeds down the taxonomy. This filters out common jargon at each step and boosts accuracy dramatically. Addition and deletion of documents are handled easily and recompilations of discriminants are done efficiently. Using the signatures, which are the function of the document and the reference node, we build a text model at each node. This model summarizes the number of documents at each node. There have been many proposals for statistical models of text generation. We will assume a Bernoulli model of document generation.

3.5.1 Bernoulli Model:

A classifier inputs a document and outputs a class. For taxonomy, the most specific class should be the output at each node. This requires a measure for each node at each level except the lower most level that classifies the document into one of the child nodes. We use discriminants for this purpose, which is defined later in this chapter. For a document with pre specified class, if the class output by the classifier does not match the prespecified class, we say the classifier misclassified that document. Typically, a classifier is trained by giving it sample documents with the class labels attached. Depending on the documents in that class, our system has a classifier at each internal node in the taxonomy, with diverse feature sets. Given a new document d, our

goal is to find a leaf node 'c' such that the probability, Pr[c/d] is maximized among all leaves thus building a multilevel classifier.

The Bernoulli model, named after James Bernoulli, is one of the simplest yet most important random processes in probability. Essentially, the process is the mathematical abstraction of coin tossing that satisfies the following assumptions.

- Each trial has two possible outcomes, generally called success and failure.
- The trials are independent. Intuitively, the outcome of one trial has no influence over the outcome of another.
- On each trial, the probability of success is p and the probability of failure is 1-p.

In this model, first picking a class generates a document d. Each class c has an associated multifaceted coin; each face represents a term t and has some associated probability $\varnothing(c,t)$ of turning up when 'tossed'. Conceptually, as the training text is being scanned, our classifier database will be organized as a very sparse three-dimensional table. One axis is for terms; the second axis is for documents and the third axis is for classes or topics. Topics have a hierarchy defined on them; here we have a tree hierarchy. The measure maintained along these dimensions (t,d,c) is called n(t,d,c), which is the number of times t occurs in $d \in c$. This number is non-zero only when $t \in d \in c$. Here, $t \in d$ means term t occurs in document d, and $d \in c$ means d is a sample document in the training set for class c.

Assuming the Bernoulli model, if document d is from class c, then, probability that the document belongs to a class is given by,

$$\mathbf{Pr[d/c] = n(d)\, C_{wt(d,t)} \sum_t f(c,t)^{wt(d,t)}}$$

Eqn (3.3)

Where, $n(d)\, C_{wt(d,t)} = n(d) / \sum_{t \in d} wt(d,t)$

n(d) - the total number of terms in the document, which is 30.

wt(d,t) - the weight of the term in the document .

f(c,t) - the number of occurrences of t in the class.

3.6 Summary:

Good taxonomies, based on the use of the classification result in more efficient information retrieval. This ensures better information retrieval and less user frustration. Depending on the corpus, we select the feature terms at each level in the hierarchy. Hence our system has a classifier at each node in the taxonomy with diverse feature sets. These feature sets are context sensitive in nature and hence help to build a multilevel classifier. Fisher's discriminant method can be used to compute the ratio of the so called between class to within class scatter. Thereby, classification of the documents based on the context is very accurate

.

CHAPTER 4

Implementation And Results

4.1 Introduction:

We have implemented our system for a few domains using JAVA. We present here the results obtained in 3 different domains Cancer, Sports and Health. The operational steps are as follows:

4.2 Algorithm:

We present here the algorithm we used in our implementation.

4.2.1 Training:

1. For each domain, we preclassified a set of training documents into a hierarchy.
2. The stop words like 'is','it','are', etc., are filtered. The stop words are specific to a particular domain.
3. Assigned different weights to the words depending on its position as title, subtitle and depending on the font size etc.,
4. Using step 3, calculated the weight of the words in the document.
5. For each document, 30 most important words have been selected to represent the document using eqn(3.1).
6. The fisher index of each term as in eqn (3.2) is calculated.
7. Sorted the terms according to decreasing Fisher index.
8. Only the top 80 to 100 discriminating terms are considered for further classification.

4.2.2 Testing:

During testing, a new document is taken up and first its 30 top terms are extracted. Then we calculated the probabilities of the document belonging to the different classes in the hierarchy. The class with the highest probability is assigned as

the category of the document. In this chapter, we summarize the hierarchies that are learnt in the various domains and also present the efficiency of the system in terms of accuracy of prediction of both training documents and test documents.

We conduct another study to look at the precision of retrieval results obtained by the system. Precision is defined as in Eqn(2.2). We have tested mostly using 2 or 3 word search queries and the repository is assumed to consist of the 70 documents that were used for training.

4.3 Cancer Domain:

First, we will consider the Cancer domain. . Some of the feature terms in the domain at each level are shown in Table 1. The hierarchy that we used for this domain for the two training sets is shown in figure(2) and figure(3) respectively. The figure shows three levels in classification. We have considered two different training sets in this domain. And for testing purpose, the first set was tested against the second and vice versa. Each training set consists of 70 documents. The training sets are shown in Appendix 1 and Appendix 2.

Different weights are assigned to the words depending on whether it is a title word, keyword, bold face etc., 30 words per documents were selected as in eqn(3.1). The Fisher index was calculated for each word using eqn(3.2). At each and every node, we cannot consider all the words clustered. This may lead to overfitting. Hence, we ourselves restricted to use only top-level Fisher index terms Using eqn(3.3), the documents are classified.

4.3.1 Stop Words in the Domain:

The stop words are filtered using the filter function. Some of the stop words that are filtered in the Cancer domain are listed below.

year	open	window	glossary	popup	and
onfocus	guide	years	women	men	some
is	was	also	upon	although	through
that	may	center	content	two	amount
width	announcement	images	pause	other	green
are	there	under	top	below	bottom

4.3.2 Results On Classification:

To test the learned hierarchy, we have given documents as input and using Eqn (3.3), we try to find the most probable class for this document. Misclassification is defined as a situation where the document does not get classified to its correct class. We have considered two types of misclassifications for each domain. The number of misclassifications when the training documents themselves are input to the system and also the number of misclassifications when new documents are fed to the system. The first kind of misclassification is called training error, whereas the second kind of misclassification is called testing error. For each domain, we call a classification correct, if the class returned by our system is the same as that specified by Google.

Training error gives the percentage of documents that are misclassified, on which the system is being trained. The following discussions gives the training and testing errors.

4.3.2.1 Training Error (Training Set 1):

Total number of Training documents	:	70
Total number of documents misclassified	:	12
Classification accuracy for training documents	:	82.85%
Training Set Error	:	17.15%

Misclassification Types:

Misclassifications at the Root node	:	Nil
Misclassifications at Cancer Types node	:	5
Misclassifications at Research node	:	7
Misclassifications at General node	:	Nil

4.3.2.2 Study on Misclassifications:

The misclassified documents at Types and Research node are classified under the General category. This is because, even though they are categorized under types or research nodes by Google, we have found that most of the contents in these documents discusses in general regarding cancer.

4.3.2.3 Training Error (Training Set 2):

Now, we discuss the results obtained with Training Set 2.

Total number of training documents	:	70
Total number of documents misclassified	:	13
Classification accuracy for training documents	:	81.43%
Training Set Error	:	18.57 %

Misclassification Types:

Misclassifications at the Root node	:	Nil
Misclassifications at Cancer Types node	:	5
Misclassifications at Research node	:	8
Misclassifications at General node	:	Nil

4.3.2.4 Study on Misclassifications:

Once again, we describe the study on misclassifications for the training set 2. Some of the documents classified under Types are misclassified under Research and some of the documents classified under Research are misclassified under Types. This is because, in the documents listed under Types of Cancer most of discussion focuses on the prevention and medical care of these cancer types. Also some of the misclassified documents at Types and Research node are classified under the General category. This is because, eventhough they are categorized under types or research nodes, most of the contents in these documents discusses in general regarding cancer.

4.3.2.5 Testing error (Test 1):

Testing error gives the percentage of new documents that are misclassified, on which the system is being tested. This test was done by testing the first set against the second. The correct class is assumed to be the one that Google gives.

Total number of testing documents	:	70
Total number of documents misclassified	:	12
Classification accuracy for testing documents	:	82.85%
Testing Set Eor	:	17.15%

Misclassifications:

Misclassifications at the root node	:	Nil
Misclassifications at Cancer Types node	:	4
Misclassifications at Research node	:	8
Misclassifications at General node	:	Nil

4.3.2.6 Study on Misclassifications:

Some of the misclassified documents at Types and Research node are classified under the General category since, most of the contents in these documents discusses in general regarding Cancer. Some of the documents classified under Types are misclassified under Research. This is because, in the documents listed under Types of Cancer most of discussion focuses on the prevention and medical care of these cancer types.

4.3.2.7 Testing error (Test 2) :

Now, we test the second hierarchy using the documents which were not used for training. Testing error gives the percentage of documents that are misclassified, on which the system is being tested.

Total number of testing documents	:	70
Total number of documents misclassified	:	12
Classification accuracy for training documents	:	82.85%
Testing Set Error	:	17.15%

Misclassifications:

Misclassifications at the root node	:	Nil
Misclassifications at Cancer Types node	:	5
Misclassifications at Research node	:	7
Misclassifications at General node	:	Nil

4.3.2.8 Study on Misclassifications:

Some of the misclassified documents at Types and Research node are classified under the General category since, most of the contents in these documents discusses in general regarding Cancer. Some of the documents classified under Types are misclassified under Research. This is because, in the documents listed under Types of Cancer most of discussion focuses on the prevention and medical care of these cancer types. Some of the documents in Types and Research are classified under general category. It is because, the contents of these documents discusses in general about Cancer.

4.3.3 Test On Retrieval:

The search engines are mainly used for relevant information retrieval, and the user's query is mainly based on the context sensitive keywords. Current techniques used in search engines are mainly based on presence or absence of these keywords. Hence, we tested on the presence of these keywords that might be commonly used by the search engine users. Using phrases for search and classification can potentially boost accuracy. The usual way to find phrases is to test a set of terms for occurrence rate far above that predicted by assuming independence between terms. We restricted our test to two term associations on the documents. The results of our observations based on recall and precision are also listed below.

4.3.3.1 Training Set 1:

Test 1:

Search query : ***Breast Cancer Treatment***

Number of documents with more than 2 term associations : 24

Actual Good Docuements : 21

Precision : 21/24 = 87.5%

Test 2:

Search query : ***Breast Mammogram Chemotherapy***

Number of documents with more than 2 term associations : 7

Actual Good Docuements : 5

Precision : 5/7 = 71.42%

Test 3:

Search query : ***Bone Cancer Treatment***

Number of documents with more than 2 term associations : 7

Actual Good Docuements : 6

Precision : 6/7 = 85.71%

Test 4:

Search query : ***Lung Cancer Treatment***

Number of documents with more than 2 term associations : 8

Actual Good Docuements : 6

Precision : 6/8 = 75%

Test 5:

Search query : ***Skin Cancer Treatment***

Number of documents with more than 2 term associations : 17

Actual Good Docuements : 14

Precision : 14/17 = 82.35%

Test 6:

Search query : ***Kidney Cancer Treatment***

Number of documents with more than 2 term associations : 7

Actual Good Docuements : 5

Precision : 5/7 = 71.42%

4.3.3.2 Justification on Testing:

The documents that are not having more than 2 term associations are found to be very small in size. They hardly have less than 90 words.

4.3.3.3 Training Set 2:

Now, We shall see the results obtained on retrieval on the training Set 2

Test 1:

Search query : ***Breast Cancer Treatment***

Number of documents with more than 2 term associations : 25

Actual Good Docuements : 21

Precision : 21/25 = 84%

Test 2:

Search query : ***Breast Mammogram Chemotherapy***

Number of documents with more than 2 term associations : 8

Actual Good Docuements : 4

Precision : 4/8 = 50%

Test 3:

Search query : ***Bone Cancer Treatment***

Number of documents with more than 2 term associations : 7

Actual Good Docuements : 5

Precision : 5/7 = 71.42%

Test 4:

Search query : ***Lung Cancer Treatment***

Number of documents with more than 2 term associations : 6

Actual Good Docuements : 4

Precision : 4/6 = 75%

Test 5:

Search query : ***Skin Cancer Treatment***

Number of documents with more than 2 term associations : 14

Actual Good Docuements : 9

Precision : 9/14 = 64.28%

Test 6:

Search query : ***Kidney Cancer Treatment***

Number of documents with more than 2 term associations : 7

Actual Good Docuements : 4

Precision : 4/7 = 57.14%

4.3.3.4 Justification on Testing:

The documents that are not having more than 2 term associations are found to be very small in size. They hardly have less than 90 words.

4.4 Sports Domain:

Next, we will consider the Sports domain. The hierarchy that we used for this domain is shown in figure 4. The figure shows three levels in classification. We have considered two different sets for training and testing in this domain. The training set consists of 72 documents and the testing set consists of 66 documents. The training set is shown in Appendix 3 and the testing set is shown in Appendix 4.

Different weights are assigned to the words depending on whether it is a title word, keyword, bold face etc., 30 words per documents were selected as in eqn(3.1). The Fisher index was calculated for each word using eqn(3.2). At each and every node, we cannot consider all the words clustered. This may lead to overfitting. Hence, we ourselves restricted to use only top-level Fisher index terms. Some of the feature terms in the domain at each level are shown in Table 4. Using eqn(3.3), the documents are classified.

4.4.1 Stop words in the domain:

The stop words are filtered using the filter function. Some of the stop words filtered in Sports domain are listed below.

upon	under	using	language	would	images	products
email	press	express	great	region	copy	press

is	was	also	upon	although	through	quit
index	size	middle	being	safety	window	span
pannel	center	file	size	middle	span	pause
are	there	under	top	below	bottom	two

4.4.2 Results On Classification:

Training error gives the percentage of documents that are misclassified, on which the system is being trained. Training and testing were done on two different sets of documents and the following discussions give the training and testing errors.

4.4.2.1 Training Error :

Now, we discuss the results obtained with Training Set.

Total number of training documents	:	72
Total number of documents misclassified	:	8
Classification accuracy for training documents	:	88.89%
Training Set Error	:	11.11%

Misclassification Types:

Misclassifications at the root node	:	Nil
Misclassifications at Ball Games	:	7
Misclassifications at Board Games	:	1
Misclassifications at Water Sports	:	Nil
Misclassifications at Motor Sports	:	Nil

4.4.2.2 Study on misclassifications:

On examining the misclassifications, under board games node, the documents are misclassified among themselves. That is, some of the cricket documents are misclassified under tennis and the tennis documents are misclassified under cricket. It is because, these documents are having most terms in common like, ball, ground, match, players, match etc., Since the occurrence of these terms are

more than the specific feature terms of the cricket or tennis node, these documents are misclassified. Other documents were too small that, they couldn't be classified.

4.4.2.3 Testing Error :

Total number of testing documents	:	66
Total number of documents misclassified	:	8
Classification accuracy for training documents	:	87.88%
Testing Set Error	:	12.12%

Misclassification Types:

Misclassifications at the root node	:	Nil
Misclassifications at Ball Games	:	4
Misclassifications at Board Game	:	1
Misclassifications at Water Sports	:	3
Misclassifications at Motor Sports	:	1

4.4.2.4 Study on Misclassifications:

While looking at the misclassifications, some of the cricket documents are misclassified under tennis and the tennis documents are misclassified under cricket. It is because, some common terms like ball, ground, match, players, match etc., are occurring in common in these documents. Hence, these documents are misclassified. Other documents were too small that, they couldn't be classified.

4.4.3 Test on Retrieval:

As explained earlier in the Cancer domain, we conducted the Test on Retrieval in the Sports domain also. The results of our observations based on recall and precision are listed below.

4.4.3.1 Training Set:

Test 1:

Search query : ***Wimbledon Tennis Match***

Number of documents with more than 2 term associations : 16

Actual Good Docuements : 13

Precision : 13/16 = 81.25%

Test 2:

Search query : ***Tennis Match Steffy***

Number of documents with more than 2 term associations : 12

Actual Good Docuements : 10

Precision : 10/12 = 83.33%

Test 3:

Search query : ***Sachin Cricket Match***

Number of documents with more than 2 term associations : 15

Actual Good Docuements : 13

Precision : 13/15 = 86.67%

Test 3:

Search query : ***Sachin Cricket Player***

Number of documents with more than 2 term associations : 14

Actual Good Docuements : 12

Precision : 12/14 = 85.71%

Test 4:

Search query : ***Swimming Diving Sports***

Number of documents with more than 2 term associations : 7

Actual Good Docuements : 5

Precision : 5/7 = 71.43%

Test 5:

Search query : ***Chess Match Player***

Number of documents with more than 2 term associations : 9

Actual Good Docuements : 6

Precision : 10/12 = 66.66%

Test 6:

Search query : ***Motor Car Race***

Number of documents with more than 2 term associations : 6

Actual Good Docuements : 5

Precision : 5/6 = 83.33%

4.4.3.2 Justification:

The documents that are not having more than 2 term associations are found to be very small in size. They hardly have less than 120 words. Some documents confer about a meticulous player and his personal life.

4.5 Health Domain:

Next, we will consider the Health domain. The hierarchy that we used for this domain is shown in figure 5. The figure shows three levels in classification. We have considered two different sets for training and testing in this domain. The training set consists of 73 documents and the testing set consists of 53 documents. The training set is shown in Appendix 5 and the testing set is shown in Appendix 6.

Different weights are assigned to the words depending on whether it is a title word, keyword, bold face etc., 30 words per documents were selected as in eqn(3.1). The Fisher index was calculated for each word using eqn(3.2). At each and every node, we cannot consider all the words clustered. This may lead to overfitting. Hence, we ourselves restricted to use only top-level Fisher index terms. Some of the feature terms in the domain at each level are shown in Table 3. Using eqn(3.3), the documents are classified.

4.5.1 Stop words in the domain:

The stop words are filtered using the filter function. Some of the stop words filtered in this domain are listed below.

committee	countries	national	quality	some	sports
techniques	approach	should	would	need	system
is	was	are	were	where	and
above	around	an	at	below	even

down	through	showed	controlled	school	animals
considered	some	management	association	centers	department

4.5.2 Results On Classification:

Training error gives the percentage of documents that are misclassified, on which the system was trained. The following discussions give the training and testing errors.

4.5.2.1 Training Error :

Total number of training documents	:	73
Total number of documents misclassified	:	16
Classification accuracy for training documents	:	78.1%
Training Set Error	:	21.9%

Misclassifications:

Misclassifications at the Root node	:	Nil
Misclassifications at General	:	10
Misclassifications at Diseases	:	4
Misclassifications at Care on health	:	2

4.5.2.2 Study on misclassifications:

The documents under General are misclassified under diseases or care on health. Since, these documents contains terms that are related with the diseases, that is, they talk about diseases and also about the health care.

4.5.2.3 Testing Error :

Total number of testing documents	:	53
Total number of documents misclassified	:	14
Classification accuracy for training documents	:	73.59%
Testing Set Error	:	26.41%

Misclassifications:

Misclassifications at the root node	:	Nil
Misclassifications at General	:	8
Misclassifications at Diseases	:	5
Misclassifications at Care on health	:	1

4.5.2.4 Study on Misclassifications:

The documents in 'General' node are misclassified under disease or care on health, since they are the general documents on health discussing about the diseases and the care.

4.5.3 Test On Retrieval:

4.5.3.1 Training Set:

Test 1:

Search query : ***doctor health query***

Number of documents with more than 2 term associations : 15

Actual Good Docuements : 12

Precision : 12/15 = 80%

Test 2:

Search query : ***diabetes insulin urination***

Number of documents with more than 2 term associations : 17

Actual Good Docuements : 14

Precision : 14/17 = 82.35%

Test 3:

Search query : ***heart attack echocardiogram***

Number of documents with more than 2 term associations : 15

Actual Good Docuements : 11

Precision : 11/15 = 73.33%

Test 4:

Search query : ***milk newborn breastfeed***

Number of documents with more than 2 term associations : 11

Actual Good Docuements : 8

Precision : 8/11 = 72.73%

Test 5:

Search query : ***cancer symptoms treatment***

Number of documents with more than 2 term associations : 15

Actual Good Docuements : 12

Precision : 12/15 = 80%

Test 6:

Search query : ***blood vessel blockage***

Number of documents with more than 2 term associations : 8

Actual Good Docuements : 5

Precision : 5/8 = 62.5%

Test 7:

Search query : ***elder citizen healthcare***

Number of documents with more than 2 term associations : 6

Actual Good Docuements : 4

Precision : 4/6 = 66.67%

4.5.3.2 Justification on Testing:

The documents that are not having more than 2 term associations are found to be very small in size. They hardly have less than 80 words.

CHAPTER 5

Conclusion And Scope Of Future Work

5.1 Conclusion:

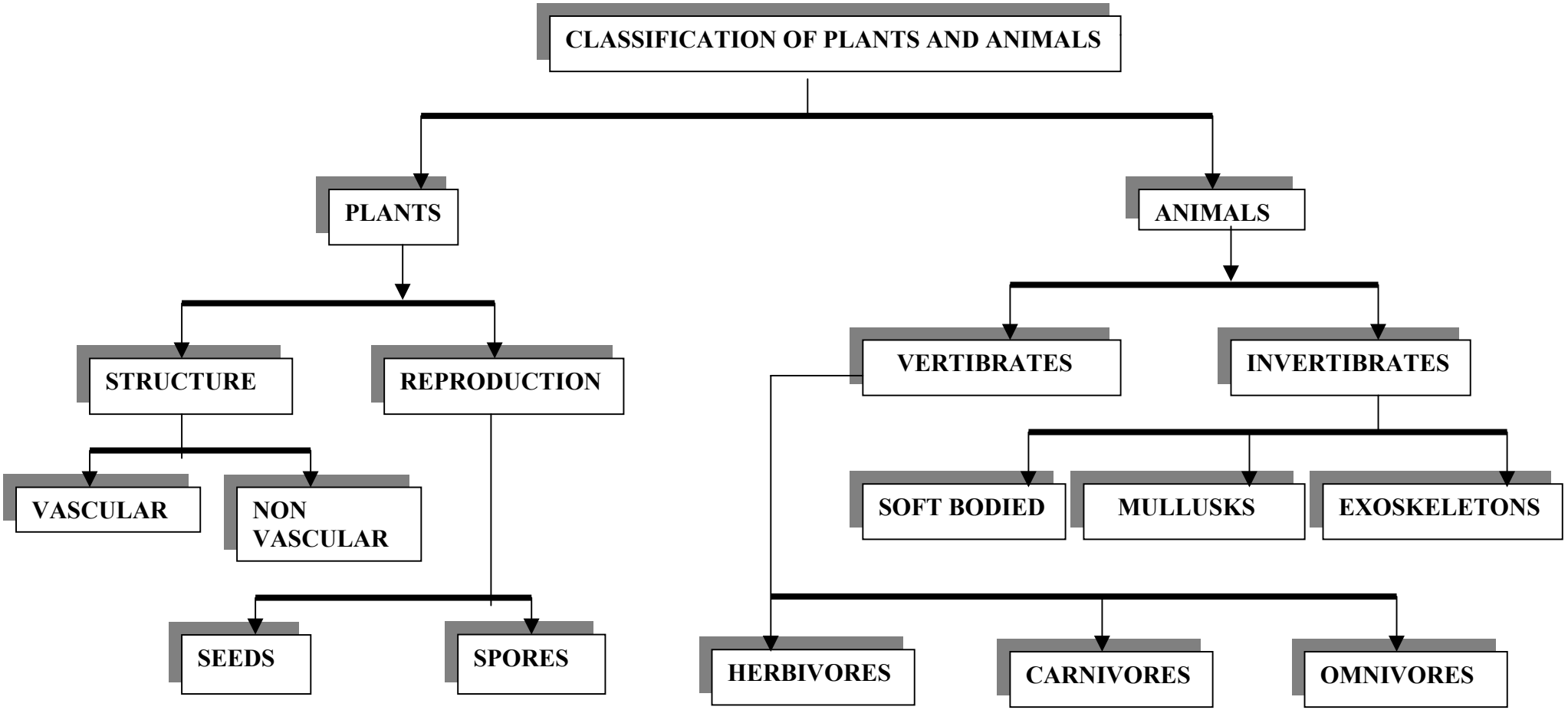
As the internet grows in size, the need for efficiently organizing the data hierarchically is on the use. We need an efficient and accurate hierarchical classification to achieve this. We suggested a system that has a context sensitive classifier at each node of the hierarchy. The terms are selected based on the Fisher's merit measure. At each node the terms receiving higher fisher index are the ones that are more discriminating analogous to that node. On the basis of the occurrence of these terms, any new document can be classified to one of the child nodes.

Results are much better if the documents follow a proper structure, as more relevant feature terms are found. More the number of documents and more even is the distribution of documents in classes, better are the results.

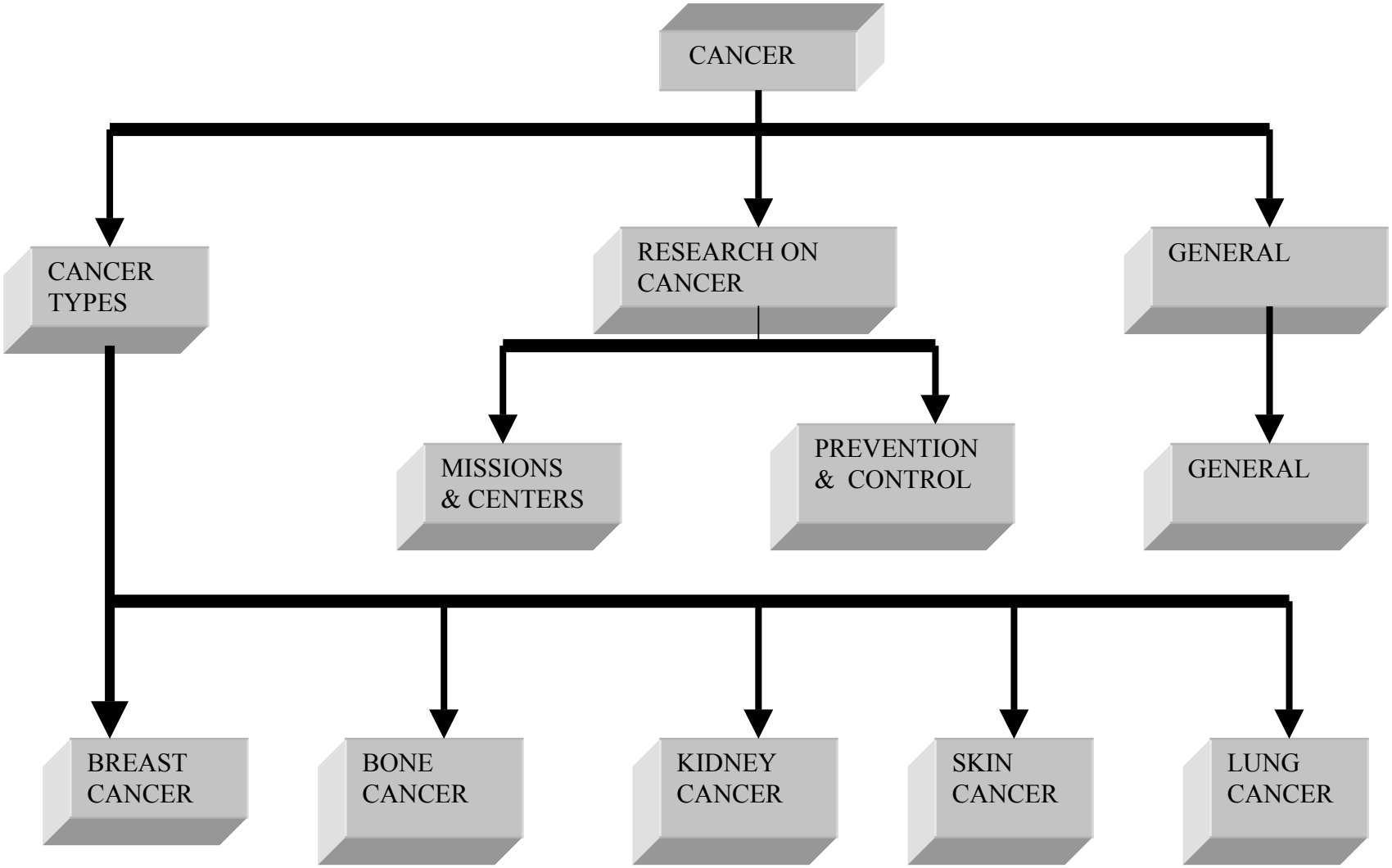
5.3 Future Work:

Whilst implementating, we assumed the independence of terms in the documents, which is far from truth. Given a term has occurred once in a document, it is more likely to occur again as compared to a term about which we have no idea. So, the work can be extended to calculate fisher value for a term taking in to account these interdependencies and correlations. Furthermore, after finding the term associations using context sensitive signature approach, we can implement search using phrases. Besides, at present we have a predefined hierarchy. Giving the system the power to introduce new nodes in the hierarchy wherever it finds necessary can do extension at this point.

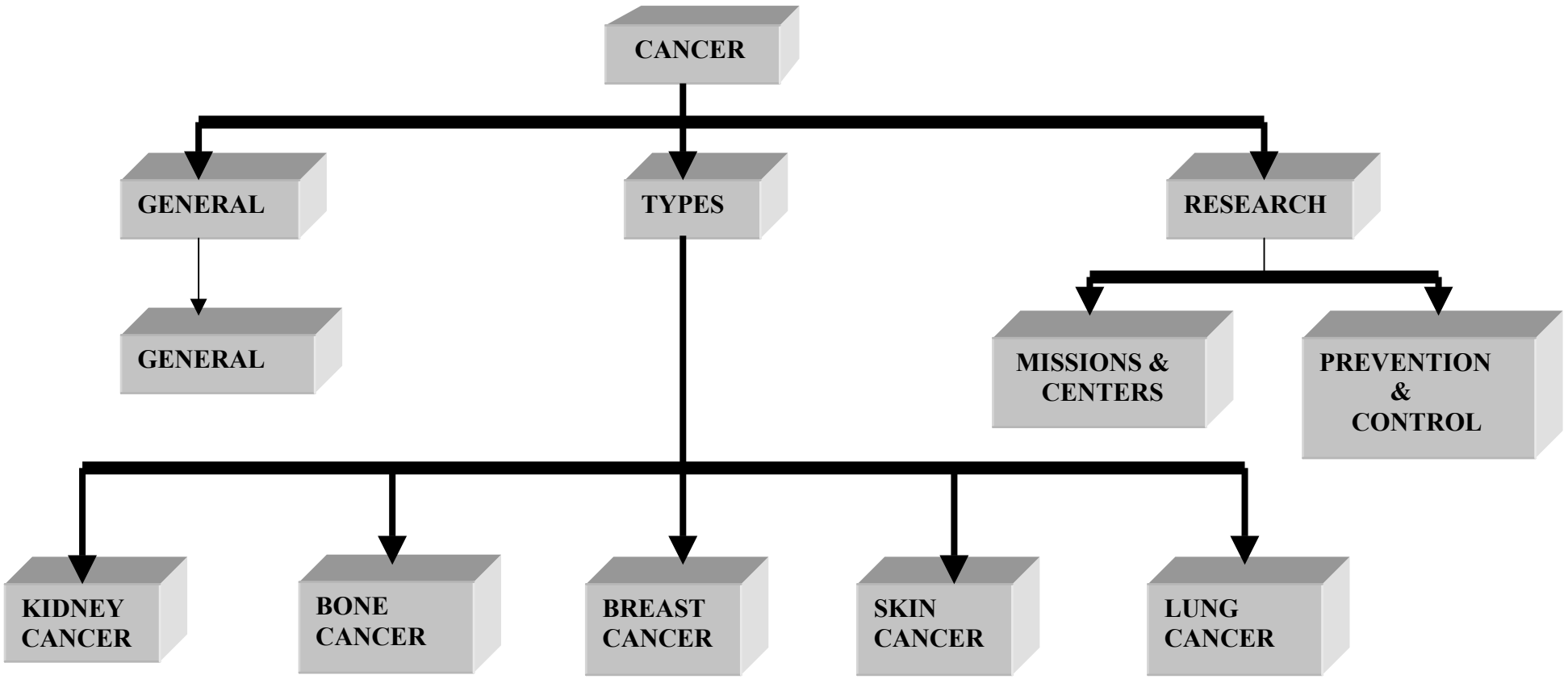
BASIC HIERARCHICAL STRUCTURE - Example (Figure 1)



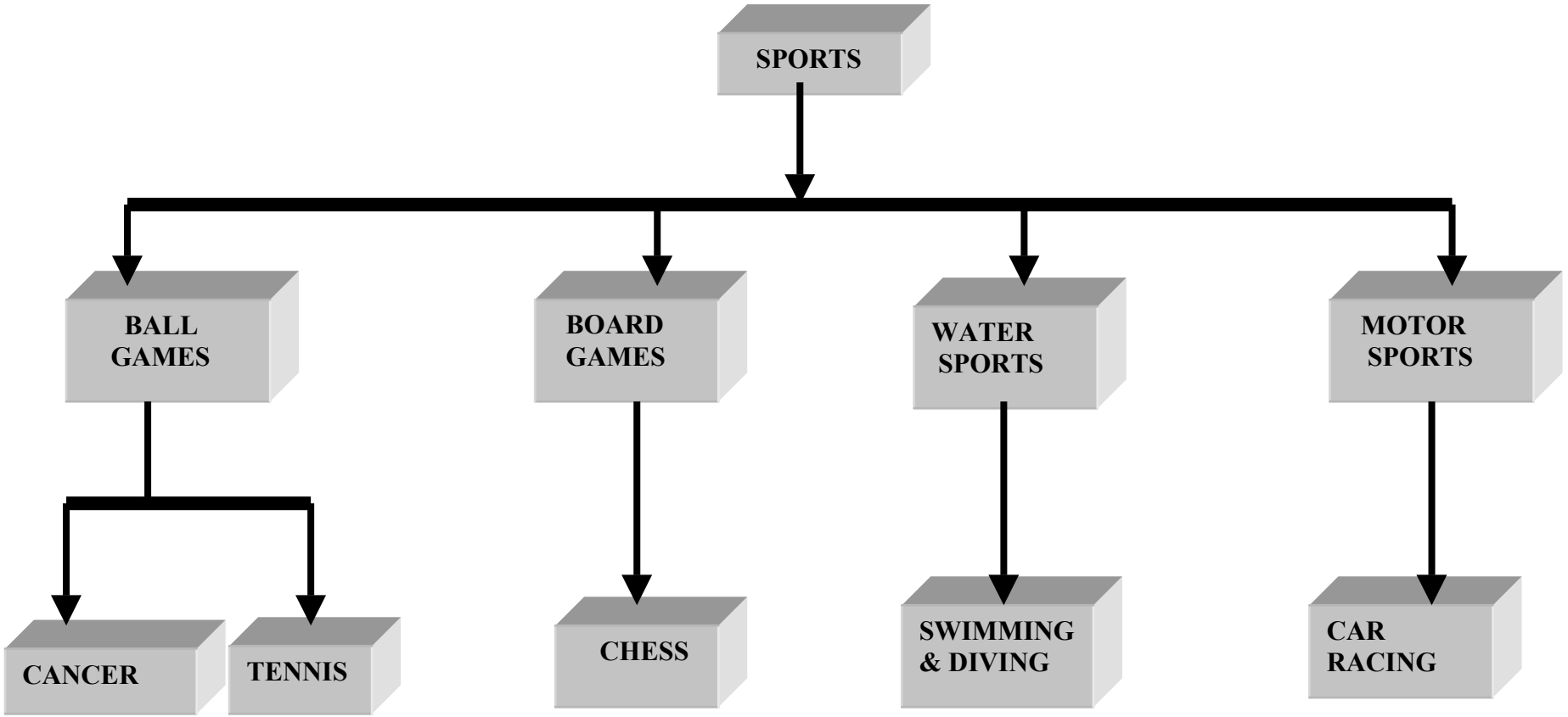
CANCER HIERARCHICAL STRUCTURE (Figure 2)



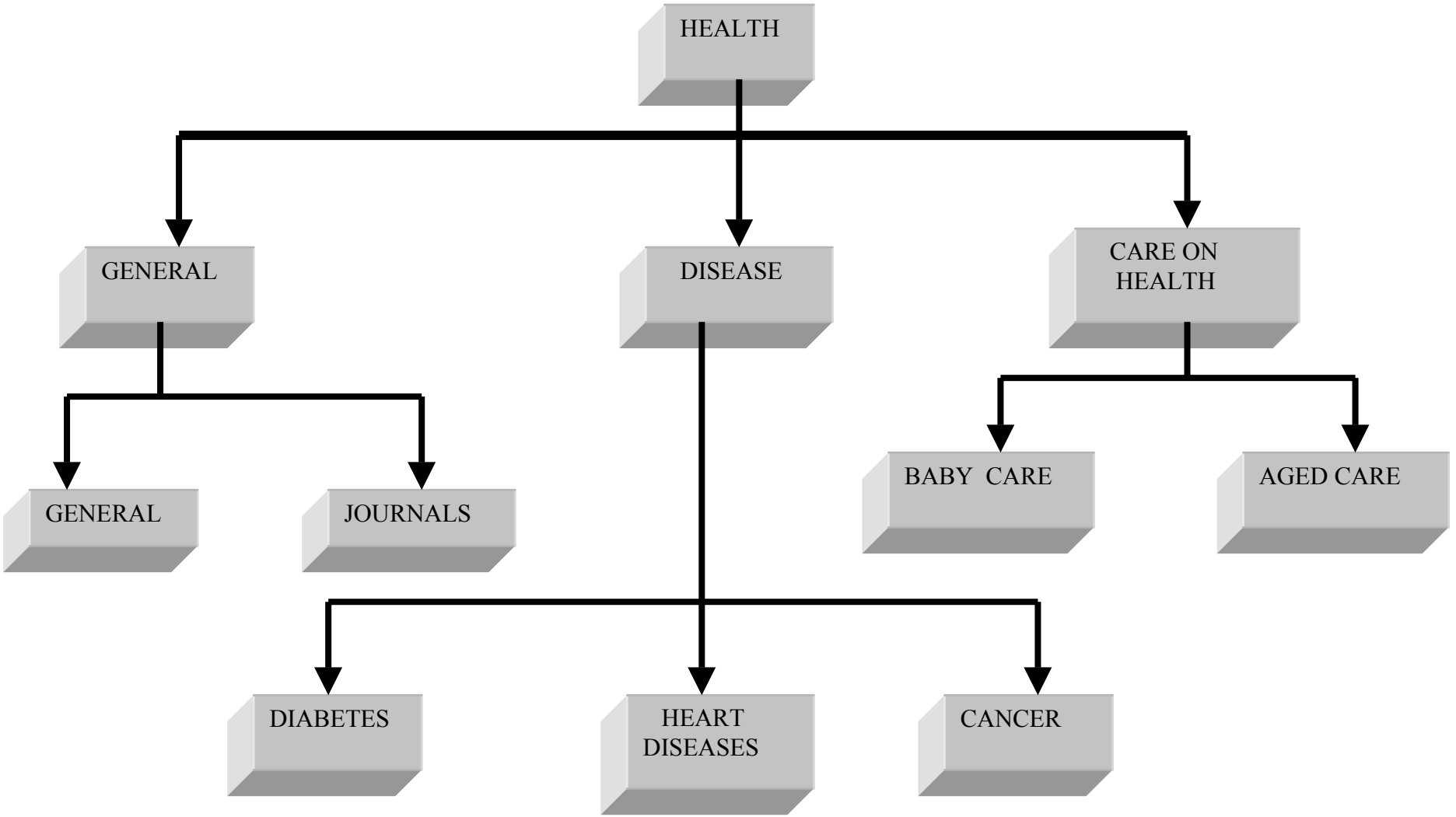
CANCER HIERARCHICAL STRUCTURE (Figure 3)



SPORTS HIERARCHICAL STRUCTURE (Figure 4)



HEALTH HIERARCHICAL STRUCTURE (Figure 5)



FEATURE TERMS IN CANCER DOMAIN

ROOT LEVEL CANCER	LEVEL 2 CANCER TYPES	LEVEL 2 RESEARCH ON CANCER	LEVEL2 GENERAL	KIDNEY CANCER	LUNG CANCER	BREAST CANCER	SKIN CANCER	BONE CANCER	MISSIONS & CENTERS	PREV. & CONTROL
Lung cancer	Nephrectomy	Researchers	Chromosome	Ultrasound	Carcinoma	Mastectomy	Melanoma	Neoplasms	Antioxidants	Metabolism
Bone cancer	Melanoma	Therapies	Syndrome	Transitional	Squamous	Tamoxifen	Malignant	Osteosarcoma	Hospitals	DNA
Radiation	Malignant	Estrogen	Viruses	Hypernephroma	Oatcell	Removal	Melanin	Histiocytoma	Doctors	Vitamins
Estrogen	Ducts	Selenium	Urology	Renalcell	Adenocarcinoma	Tissues	Melanocytes	Sarcoma	Cancerhome	Fibers
Therapies	Dermis	Healing	Biological	Malignant	Largecell	Lumpectomy	Sunburn	Marrow	Partners	Folicacid
Urologists	Adenocarcinoma	Treating	Lymphoma	Urination	Smallcell	Lymph	Mole	Bloodcell	Clinical	Selenium
Samples	Carcinoid	Investigators	Vitamins	Sugar	Pneumonectomy	Sentinal	Basalcell	Marrow	Research	Calcium
Internal	Melanin	Cancercenter	Antioxidants	Bladder	Chemotherapy	Selfexam	Carcinoma	Vein	Cancerfund	Dietery
Biological	Melanocytes	Missions	Treatment	Nephropathy	Cough	Herception	Squamous	Myeloma	Resources	Healing
Chromosome	Urological	Laboratories	Hybridization	Analgesic	Radiation	Estrogen	Rodant	Plasma	Missions	Therapies
Clone	Urination	Calcium	Radiation	Smoking	Surgical	Endometrial	Dermis	Neoplasma	Donate	Antioxidants
Hybridization	Bloodclots	Dietery	Bladder	Prostatic	Microscopy	Tamoxifen	Fairskin	Ocaltinonin	Funding	Multivitamins
Genomic	Ultrasound	Retinoids	Lymph	Urography	Radiotherapy	Radiation	Lymphoma	Puberty	Corporation	Coenzyme
Recurrent	Microscopic	Folicacid	Basalcell	Peptic	Breathing	Digital	Sarcoma	Liposarcoma	Society	Laboratory
Cavity	Surgery	Multivitamins	Therapy	Dyalisis	Haemoptysis	Angiogenesis	Metasize	degeneration	Institute	Retinoids
Facilities	Angiogenesis	Biotechnology	Bladder	Tomography	Mammography	Mammography	Cryosurgery	Cartilage	Resources	Survival
Nephrectomy	Electrodesiccation	Molecular	Ovarian	Scanning	Pulmmary	Ductal	Electrodesiccation	Osteosarcoma	Biotechnology	Immune
Transplant	Liposarcoma	Chemistry	Leukema	Magnetic	Aorta	Lavage	Curet	Chondrosarcoma	Analysis	Healing
Lymph	Incisional	Science	Melatonin	Resonance	Mesothelima	Biopsy	Dermabrasion	Ewing	Molecular	Health
Basal	Biopsy	Drug	Therapies	Ultrasound	Malignancy	Hormone	Excision	Osteoid	Technology	Menopause
Bladder	Estrogen	Nutrients	Urology	Arteriography	Asymptotic	Receptor	Prognosis	Retinoblastoma	Diagnosis	Nutrients
Corectal	Lumpectomy	Medicine	Hospitals	Uteroscopy	Resection	Ultrasound	Cutaneous	Angiogram	Collaboration	Diet
Ovarian	Selfexam	Immune	Samples	Ureter	Purgatives	Lobular	Nonmelanoma	Orthopedic	Analysis	Bloodstream
Testicular	Needle	Calcium	Facilities	Pelvis	Palliative	Breast	Lump	Needle	Chemistry	Estrogen
Aberrations	Radiation	Dietery	Missions	Scelerosis	Lung	Intraductal	Skin	Incisional	Science	Medicine
scientific	Analgesic	Antioxidants	Surgery	Kidney	Chronic	menstrual	Laser	Biopsy	Drug	Treating

Table showing the feature terms in Cancer domain(Table 1)

FEATURE TERMS IN SPORTS DOMAIN

ROOT LEVEL SPORTS	BALL GAMES	BOARD GAMES	WATER SPORTS	MOTOR SPORTS	CRICKET	TENNIS	CHESS	SWIMMING & DIVING	CAR RAC-ING
Cricketer	Wimbledon	Move	Swimming	Racing	Batsman	Martina	Move	Swimming	Racing
Tennisball	Ground	Chess	Paddlers	Racecar	Bradman	Tennisball	Chess	Paddlers	Racecar
Chessboard	Court	King	Water	Formulaone	Sachin	Usopen	King	Water	Formulaone
Swimmer	Steffy	Queen	Aquaskier	Sportscar	Gavaskar	Wimbledon	Queen	Aquaskier	Sportscar
Racecar	Martina	Board	Diving	Driving	Bowling	Hinges	Board	Diving	Driving
Sachin	Hinges	Anand	Skiing	Driver	Batting	Navratilova	Anand	Skiing	Driver
Graf	Sachin	Kasparov	Diver	Champion	Bowler	Sampras	Kasparov	Diver	Champion
Kasparov	Bowling	Player	Butterfly	Grandprix	Fielding	Becker	Player	Butterfly	Grandprix
Queen	Fielding	Pawn	Freestyle	Raceclub	Umpire	Grandslam	Pawn	Freestyle	Raceclub
King	Wicket	Vladimir	Boater	Formulathree	Allrounder	Steffy	Vladimir	Boater	Formulathree
Pawn	Umpire	Kramnik	Stroke	Track	Record	Jim	Kramnik	Stroke	Track
Diving	Ganguly	Tournament	Aquaskier	Sprint	Worldcup	Agassi	Tournament	Aquaskier	Sprint
Championship	Sampras	Chessmaster	Olympic	Cockpits	Jadeja	Juniors	Chessmaster	Olympic	Cockpits
Umpire	Usopen	Grandmaster	Breaststroke	Lightweight	Dravid	Competition	Grandmaster	Breaststroke	Lightweight
Bowling	Becker	Raymond	Dogstroke	Championship	Wicket	Daviscup	Raymond	Dogstroke	Championship
Throwing	Agassi	John	Windmill	Britishrace	Botham	Coaching	John	Windmill	Britishrace
Fielding	Williams	Watson	Backstroke	Automobile	Australia	Ranking	Watson	Backstroke	Automobile
Skiing	Navratilova	Expert	Title	Horsepower	India	Monica	Expert	Title	Horsepower
Diver	Grandslam	Opening	Championship	Frontengine	England	Peter	Opening	Championship	Frontengine

Table showing the Feature Terms in Sports domain (Table 2)

FEATURE TERMS IN HEALTH DOMAIN

RootLevel Health	General	Diseases	Care On Health	Journals	Diabetes	Heart Disease	Cancer	Baby Care	Aged Care
Nursing	Directory	Cyanosis	Nursing	Browse	Insulin	Congenital	Radiotherapy	Nursing	Mental
Radiotherapy	Assessment	Telralogy	Latching	Archive	Mellitus	Myxomatous	Mammography	Latching	Palliative
Myxomatous	Childhood	Ventricular	Pallative	Articles	Glucose	Atherosclerosis	Squamous	Clutch	Convalescent
Telralogy	Primary	Melanoma	Convalescent	Instructions	Thirst	Calcium	Melanin	Cradle	Indigenous
Kidshealth	Rural	Mellitus	Indigenous	Feedback	Tingling	Arrhythmias	Histiocytoma	Mother	Corporate
Promotion	Safety	Glucose	Breastfeed	International	Nausea	Valvular	Osteosarcoma	Breastfeed	Financial
Glucose	Promotion	Tingling	Nutrition	Archive	Blurred	Syndrome	Prognosis	Feed	Adopted
Nutrition	Epidomiology	Fatigue	Milk	Publication	Triglyceride	Cyanosis	Cryosurgery	Nutrition	Assessment
Authors	Tuberculosis	Valvular	Sucking	Schedule	Gangrene	Telralogy	Osteoid	Milk	Ageing
Publishers	Immunization	Telralogy	Bottlefeed	Contents	Amputation	Ventricular	Angiogram	Sucking	Residential
Counseling	Healthcare	Sweating	Podiatry	Tables	Gestational	Electrocardiogram	Uteroscopy	Bottlefeed	Restructuring
Breastfeed	Schedule	Coronary	Counseling	Email	Pancreas	Pulmonary	Nephropathy	Lying	Funeral
Gangrene	Mental	Osteoid	Pediatrics	Guestbook	Fatigue	Sweating	Neoplasms	Kidshealth	Counseling
Amputation	Privatized	Haemoptysis	Growth	Issue	Glucose	Dizziness	Melanoma	Bottlefeed	Podiatry
Massive	Counseling	Dermis	Pension	Authors	Gestational	Massive	Lobular	Pediatrics	Physiotherapy
Artery	Corporate	Insulin	Financial	Publishers	Steroids	Angina	Haemoptysis	Lactation	Wheelchair
Nephropathy	Financial	Blurred	Bottlefeed	Library	Diuretics	Coronary	Estrogen	Pregnancy	Hearingaid
Neoplasms	Email	Triglyceride	Mother	Guide	Genitals	Artery	Dermis	Growth	Pension
Lobular	Bureau	Melanin	Restructuring	Contents	Fasting	Cardic	Basalcell	Food	Privatized

Table showing the feature terms in Health domain (Table 3)

Appendix 1

Cancer Training Set 1

1. http://cancernet.nci.nih.gov/
2. http://www.sharedexperience.org/cancerindex.lasso
3. http://cbshealthwatch.medscape.com/cx/viewarticle/403501_2
4. http://www.ncbi.nlm.nih.gov/ncicgap/
5. http://www.cancerlynx.com/
6. http://www.cancerbacup.org.uk/
7. http://cancer.about.com/mbody.htm
8. http://www.children-cancer.com/home/index.html
9. http://www.cancereducation.com/cancersyspagesnb/a/statyb.cfm?pg=pfctr
10. http://www.cancer-drug.com/
11. http://www.cancerfacts.com/
12. http://www.cancersupportivecare.com/riskintro.html#where
13. http://www.breast-cancers.com/
14. http://www.bcdg.org/diagnosis/types.html
15. http://members.tripod.com/~Breast_Cancer101/facts.html
16. http://www.healingwell.com/library/breastcancer/lewis1.htm
17. http://www.winabc.org/newweb/breast-health/breast-anatomy-and-physiology.htm
18. http://home.earthlink.net/~tomcarla/breastcancer/cancer.htm
19. http://www.cancerhelp.com/Diagnoss.htm
20. http://www.imaginis.com/breasthealth/menu-diagnosis.asp
21. http://www.babcn.org/health.htm
22. http://www.cancer-info.com/breast.htm
23. http://www.infobreastcancer.cyberus.ca/diagbc.htm
24. http://www.associatedurologists.com/kidney.html
25. http://www.uro.com/renalca.htm
26. http://www.netdoctor.co.uk/diseases/facts/kidneycancer.htm
27. http://www.mskcc.org/mskcc/html/362.cfm
28. http://www.clevelandclinic.org/urology/patients/kidney/cancer.htm
29. http://www.skin-cancers.net/
30. http://www.maui.net/~southsky/introto.html#causes
31. http://www.melanoma.com/whatis/index.htm
32. http://www.jas.tj/skincancer/facts.html
33. http://www.aad.org/SkinCancerNews/WhatIsSkinCancer/SCancerFacts.html
34. http://www.cancerbacup.org.uk/info/skin/skin-5.htm
35. http://www.meb.uni-bonn.de/cancernet/201228.html
36. http://cancer.about.com/cs/skincancer/
37. http://www.cancerlinksusa.com/skin/treatment_pt/description.htm#5
38. http://www.cancerindex.org/clinks2s.htm
39. http://www.thecancerweb.com/lung/wynk/
40. http://hebw.uwcm.ac.uk/cancers/Chapter1.html

41. http://www.quitnow.info.au/oncology.html
42. http://www.tirgan.com/lung.htm
43. http://www.med.umich.edu/1libr/cancer/smokg04.htm
44. http://www.bocaradiology.com/Procedures/LungCA.html
45. http://www.ricancercouncil.org/facts/lungfacts.htm
46. http://www.meds.com/pdq/nonsmallcell_pat.html#5
47. http://www.cancercenter.com/home/90/lung/lung1.cfm
48. http://www.cancer.gov/prevention/
49. http://www.cancerresearchamerica.org/
50. http://www.breastcancerstrategies.com/
51. http://pinc.com/healthnews/breastcancer.html
52. http://www.cancer-options.net/html/cancer_prevention.html
53. http://www.iarc.fr/
54. http://www.ctrf.org/
55. http://www.fhcrc.org/science/
56. http://www.bccrc.ca/
57. http://web.ncifcrf.gov/
58. http://www.researchforacure.com/site/PageServer
59. http://www-dcs.nci.nih.gov/branches/
60. http://www.uihealthcare.com/depts/cancercenter/
61. http://www.asu.edu/clas/cancer_research/
62. http://ben-may.bsd.uchicago.edu/bmi/common/home.html
63. http://ccr.nci.nih.gov/
64. http://www.aicr.org/r111601.htm
65. http://www.cancergroup.com/em19.html
66. http://cis.nci.nih.gov/fact/6_26.htm
67. http://www.cancerindex.org/ccw/faq/
68. http://bonetumor.org/page33.html
69. http://www.uicc.org/
70. http://www.geocities.com/HotSprings/9041/

Appendix 2

Cancer Training Set 2

1. http://cancernet.nci.nih.gov/
2. http://www.sharedexperience.org/cancerindex.lasso
3. http://cbshealthwatch.medscape.com/cx/viewarticle/403501_2
4. http://www.ncbi.nlm.nih.gov/ncicgap/
5. http://www.cancerlynx.com/
6. http://www.cancerbacup.org.uk/
7. http://cancer.about.com/mbody.htm
8. http://www.children-cancer.com/home/index.html
9. http://www.cancereducation.com/cancersyspagesnb/a/statyb.cfm?pg=pfctr
10. http://www.cancer-drug.com/
11. http://www.cancerfacts.com/
12. http://www.cancersupportivecare.com/riskintro.html#where
13. http://www.imperialcancer.co.uk/search_results01.cfm?SecID=721&DocID=37&CatID=1
14. http://www.associatedurologists.com/kidney.html
15. http://www.mskcc.org/mskcc/html/362.cfm
16. http://www.med.umich.edu/1libr/cancer/renal02.htm#who
17. http://cancernet.nci.nih.gov/wyntk_pubs/kidney.htm#4
18. http://cis.nci.nih.gov/fact/6_26.htm
19. http://bonetumor.org/page33.html
20. http://www.uicc.org/
21. http://www.geocities.com/HotSprings/9041/
22. http://www.canceranswers.com/Bone.Cancer.html
23. http://www.medformation.com/mf/crsaa.nsf/crs/cancbon.htm
24. http://www.breast-cancers.com/
25. http://www.bcdg.org/diagnosis/types.html
26. http://www.winabc.org/newweb/breast-health/breast-anatomy-and-physiology.htm
27. http://www.cancerhelp.com/Diagnoss.htm
28. http://www.imaginis.com/breasthealth/menu-diagnosis.asp
29. http://www.babcn.org/health.htm
30. http://www.infobreastcancer.cyberus.ca/diagbc.htm
31. http://medlib.med.utah.edu/WebPath/TUTORIAL/BREAST/BREAST.html *
32. http://www.blackwomenshealth.com/breast_cancer.htm
33. http://www.breastthermography.com/what_is_breast_cancer.htm
34. http://members.ozemail.com.au/~glensan/cancer.htm
35. http://www.umm.edu/skincancer/melanoma.htm
36. http://www.cancerindex.org/clinks2s.htm
37. http://www.maui.net/~southsky/introto.html#is
38. http://www.tirgan.com/skin.htm

39. http://www.umm.edu/skincancer/basal.htm
40. http://www.nedermatology.com/skincancer/ak/index.html
41. http://www.med.umich.edu/1libr/cancer/skin04.htm
42. http://www.cancerlinksusa.com/skin/wynk/glossary.htm#nonmelanoma%20skin%20cancer
43. http://www.medformation.com/mf/crsaa.nsf/crs/skincanc.htm
44. http://www.nedermatology.com/skincancer/mm/index.html
45. http://www.bocaradiology.com/Procedures/LungCA.html
46. http://www.thecancerweb.com/lung/wynk/
47. http://www.roycastle.org/lungcancerresearch/high/what/4.html
48. http://www.cdc.gov/cancer/
49. http://www.strang.org/
50. http://visitors.healthandwellnessclub.com/mag_articles/article_8.asp
51. http://www.hcfa.gov/medicaid/bccpt/default.htm
52. http://www.pcrm.org/health/Preventive_Medicine/foods_for_cancer_prevention.html
53. http://www.cancerlinksusa.com/centers.htm
54. http://www.ctrf.org/research.htm
55. http://www.researchforacure.com/site/PageServer?pagename=research&JServSessionIdr012=hwubt1sljb.app7b
56. http://www.bcrfcure.org/
57. http://www.nbcc.org.au/pages/info/research.htm
58. http://www.cancercenter.com/home/index.cfm
59. http://www.fhcrc.org/visitor/hutch_story/
60. http://www.uchicago.edu/uchi/resteach/groups.html
61. http://www.hsph.harvard.edu/Organizations/Canprevent/links.html
62. http://ccr.nci.nih.gov/news/press/JNCI_CCR.asp
63. http://www.fhcrc.org/science/
64. http://www.asbestos-attorney.com/hospitalcenters.htm
65. http://ccr.nci.nih.gov/news/press/JNCI_CCR.asp
66. http://www.cancerresearch.org/
67. http://www.cancer.mednet.ucla.edu/newsmedia/news/pr091499.html
68. http://www.hcfa.gov/medicaid/bccpt/default.htm
69. http://pinc.com/healthnews/breastcancer.html
70. http://www.cancer-options.net/html/cancer_prevention.html

Appendix 3

Sports Training Set

1. http://www-usa.cricket.org/link_to_database/PLAYERS/AUS/B/BRADMAN_DG_02000492/ARTICLES/BRADMAN_PAGES/
2. http://www.yehhaicricket.com/matches2001/australia_india/articles/artciles.html
3. http://www.azharuddin.com/thecricketer.html
4. http://www.yehhaicricket.com/matches2001/australia_india/analysis/analysis.html
5. http://www.geocities.com/MotorCity/Flats/8438/gavaskar.htm
6. http://www.yehhaicricket.com/matches2001/australia_india/matchsummary/match_summary.html
7. http://www-usa.cricket.org/link_to_database/PLAYERS/IND/T/TENDULKAR_SR_06001934/SPECIAL/main.html
8. http://members.tripod.com/~kkv/st.htm
9. http://www.yehhaicricket.com/matches2001/australia_india/swot/team_preview.html
10. http://cricket.indiatimes.com/homepages/sachin/stats.html
11. http://members.tripod.com/~VineetV/HOMEPAGE.HTM
12. http://www.englishsport.com/cricketheaven/
13. http://www.canoe.ca/SlamCricket/home.html
14. http://www.yehhaicricket.com/india/Rahuld/rahul.html#
15. http://www.espnstar.com/jsp/cda/studio/id=5843&aid=32223&ecode='LEFTNAVBAR_COL1'&colid=23597studio_pastcoldetail.html
16. http://dmoz.org/Sports/Cricket/News_and_Media/
17. http://directory.google.com/Top/Sports/Cricket/News_and_Media/
18. http://it.uts.edu.au/news/media/010110cohen.html
19. http://tennispro.www2.50megs.com/
20. http://sports.yahoo.com/ten/news/capriati01.html
21. http://www.worldsportsmen.f2s.com/hingis.html
22. http://www.worldsportsmen.f2s.com/navratilova.html
23. http://www.worldsportsmen.f2s.com/swilliams.html
24. http://www.worldsportsmen.f2s.com/sampras.html
25. http://www.worldsportsmen.f2s.com/becker.html
26. http://www.worldsportsmen.f2s.com/novotna.html
27. http://www.worldsportsmen.f2s.com/seles.html
28. http://www.worldsportsmen.f2s.com/mcenroe.html
29. http://digilander.iol.it/ander75/grandslam.htm
30. http://www.goindiago.com/sports/tennis/tennishistory.htm
31. http://www.usopen.org/
32. http://www.tennisw.com/
33. http://www.sportsmediainc.net/tennisweek/

34. http://www.soyouwanna.com/site/syws/chess/chess.html
35. http://directory.google.com/Top/Games/Board_Games/C/Chess/News_and_Media/
36. http://dir.yahoo.com/Recreation/games/board_games/chess/
37. http://dmoz.org/Games/Board_Games/C/Chess/Software/Macintosh/
38. http://www.soyouwanna.com/site/syws/chess/chess2.html
39. http://dir.123india.com/sports/chess/
40. http://www.soyouwanna.com/site/syws/chess/chess3.html
41. http://chessmania.master.com/texis/master/search/%2B/Top/Games/Board_Games/C/Chess/Correspondence_Chess
42. http://www.soyouwanna.com/site/syws/chess/chess5.html
43. http://www.soyouwanna.com/site/syws/chess/chess4.html
44. http://www.aquaskier.com/
45. http://sports.tamu.edu/sports/wswimming/
46. http://canoekayak.about.com/library/weekly/blWhyDying.htm
47. http://www.waveclubwatersports.com/
48. http://www.ncaachampionships.com/swim/wswim/
49. http://www.auburn.edu/athletics/sd/2002outlook.html
50. http://sports.tamu.edu/sports/mswimming/
51. http://www.outdoorwatersports.com/Default.htm
52. http://directory.google.com/Top/Sports/Water_Sports/Swimming_and_Diving/
53. http://www.ncaachampionships.com/swim/mswim/index.html
54. http://www.lycos.co.uk/dir/Sports/Water_Sports/Swimming/
55. http://directory.excite.com/sports/outdoors/water_sports/swimming_and_diving/news_and_guides/
56. http://www.openhere.com/kids1/sports-and-recreation/water-sports/swimming/
57. http://kunani.com/Sports/Water_Sports/Swimming_and_Diving/
58. http://dmoz.org/Sports/Water_Sports/Water_Skiing_and_Wakeboarding/
59. http://dir.yahoo.com/Recreation/Sports/Waterskiing/
60. http://directory.google.com/Top/Sports/Water_Sports/Water_Skiing_and_Wakeboarding/
61. http://directory.excite.com/sports/water_sports/water_skiing/
62. http://www.californiadelta.org/waterski.htm
63. http://www.pacesupercross.com/
64. http://www.softcom.net/users/kartatck/
65. http://www.na-motorsports.com/
66. http://www.nascar.com/
67. http://directory.excite.com/sports/motor_sports
68. http://www.ukmotorsport.com/racmsa/starting/car_racing.html
69. http://directory.excite.com/sports/sports_a_z/motor_sports/other_classes/
70. http://www.mulsannescorner.com/history.htm
71. http://www.lycos.co.uk/dir/Sports/Motor_Sports/
72. http://www.lycos.co.uk/dir/Sports/Motor_Sports/

Appendix 4

Sports Test Set

1. http://www-usa.cricket.org/link_to_database/PLAYERS/AUS/B/BRADMAN_DG_02000492/ARTICLES/BRADMAN_PAGES/
2. http://www.yehhaicricket.com/matches2001/australia_india/articles/artciles.html
3. http://www.yehhaicricket.com/matches2001/australia_india/analysis/analysis.html
4. http://www.yehhaicricket.com/matches2001/australia_india/matchsummary/match_summary.html
5. http://members.tripod.com/~kkv/st.htm
6. http://www.yehhaicricket.com/matches2001/australia_india/swot/team_preview.html
7. http://cricket.indiatimes.com/homepages/sachin/stats.html
8. http://www.englishsport.com/cricketheaven/
9. http://www.worldsportsmen.f2s.com/navratilova.html
10. http://www.yehhaicricket.com/india/Rahuld/rahul.html#
11. http://www.cricket365.com/
12. http://www.cricketworld.com/
13. http://www.cricket-online.org/news/archive/2001/December/07_DEC_2001_ASHETH.html
14. http://www.cricmania.com/
15. http://dmoz.org/Sports/Cricket/News_and_Media/
16. http://directory.google.com/Top/Sports/Cricket/News_and_Media/
17. http://tennispro.www2.50megs.com/
18. http://sports.yahoo.com/ten/news/capriati01.html
19. http://www.worldsportsmen.f2s.com/hingis.html
20. http://www.worldsportsmen.f2s.com/navratilova.html
21. http://www.worldsportsmen.f2s.com/swilliams.html
22. http://www.worldsportsmen.f2s.com/sampras.html
23. http://www.worldsportsmen.f2s.com/novotna.html
24. http://www.worldsportsmen.f2s.com/seles.html
25. http://www.worldsportsmen.f2s.com/mcenroe.html
26. http://www.hope.edu/pr/athletics/womenswim/main.htm
27. http://www.goindiago.com/sports/tennis/tennishistory.htm
28. http://www.usopen.org/
29. http://news.bbc.co.uk/sport/hi/english/tennis/default.stm
30. http://archive.sportserver.com/newsroom/sports/oth/1998/oth/ten/feat/archive/052298/ten71273.html
31. http://archive.sportserver.com/newsroom/ap/oth/1998/oth/ten/feat/archive/052298/ten50011.html
32. http://www.chesscenter.com/
33. http://sports.tamu.edu/sports/mswimming/
34. http://www.gmchess.com/
35. http://www.internetchess.com/

36. http://dmoz.org/Games/Board_Games/C/Chess/Software/Macintosh/
37. http://www.soyouwanna.com/site/syws/chess/chess2.html
38. http://dir.123india.com/sports/chess/
39. http://www.soyouwanna.com/site/syws/chess/chess3.html
40. http://chessmania.master.com/texis/master/search/%2B/Top/Games/Board_Games/C/Chess/Correspondence_Chess
41. http://www.soyouwanna.com/site/syws/chess/chess5.html
42. http://edschool.csuhayward.edu/departments/kpe/I.C.Sports/swim-women.html
43. http://www.hope.edu/pr/athletics/womenswim/main.htm
44. http://www.aquaskier.com/
45. http://sports.tamu.edu/sports/wswimming/
46. http://canoekayak.about.com/library/weekly/blWhyDying.htm
47. http://www.waveclubwatersports.com/
48. http://www.ncaachampionships.com/swim/wswim/
49. http://www.auburn.edu/athletics/sd/2002outlook.html
50. http://sports.tamu.edu/sports/mswimming/
51. http://www.outdoorwatersports.com/Default.htm
52. http://directory.google.com/Top/Sports/Water_Sports/Swimming_and_Diving/
53. http://www.ncaachampionships.com/swim/mswim/index.html
54. http://www.lycos.co.uk/dir/Sports/Water_Sports/Swimming/
55. http://sports.yahoo.com/ten/news/capriati01.html
56. http://www.openhere.com/kids1/sports-and-recreation/water-sports/swimming/
57. http://kunani.com/Sports/Water_Sports/Swimming_and_Diving/
58. http://www.motorsport.com/
59. http://www.mitsubishi-motors.co.jp/motorsports/
60. http://www.pacesupercross.com/
61. http://www.softcom.net/users/kartatck/
62. http://www.na-motorsports.com/
63. http://www.nascar.com/
64. http://directory.excite.com/sports/motor_sports
65. http://sportsillustrated.cnn.com/motorsports/world/news/2001/04/15/f1_san_marino/
66. http://www.ukmotorsport.com/racmsa/starting/car_racing.html

Appendix 5

Health Training Set

1. http://www.cdc.gov/travel/
2. http://www.hc-sc.gc.ca/english/care/index.html
3. http://www.acsh.org/
4. http://www.bbc.co.uk/health/
5. http://www.doh.state.fl.us/
6. http://www.intelihealth.com/IH/ihtIH/WSIHW000/408/408.html
7. http://www.state.tn.us/health/
8. http://www.canoe.ca/Health/home.html
9. http://www.kdhe.state.ks.us/health/index.html
10. http://www.fhi.org/
11. http://www.health.qld.gov.au/HealthyLiving/default.htm
12. http://www.lib.uiowa.edu/hardin/md/ej.html
13. http://www.blackwell-science.com/uk/journals.htm
14. http://www.freemedicaljournals.com/
15. http://www.library.adelaide.edu.au/guide/med/menthealth/jnl.html
16. http://www.aje.oupjournals.org/
17. http://www.springerjournals.com/catalog.htm
18. http://www.diabetic.org.uk/
19. http://www.life-with-diabetes.com/
20. http://www.viahealth.org/disease/diabetes/type1.htm
21. http://www.lillydiabetes.com/Education/DiabetesTypes.cfm
22. http://www.viahealth.org/disease/diabetes/gesta.htm
23. http://www.viahealth.org/disease/diabetes/type2.htm
24. http://www.netdoctor.co.uk/diseases/facts/diabetes.htm
25. http://www.mamashealth.com/Diabetes.asp
26. http://www.diabetes.ca/about_diabetes/thefacts.html
27. http://www.healthinsite.gov.au/T.cfm?PID=1544
28. http://my.webmd.com/content/dmk/dmk_article_5963070
29. http://hlunix.hl.state.ut.us/cfhs/chronic/diabetes/New_Folder/types_of_diabetes.htm
30. http://www.niddk.nih.gov/health/diabetes/pubs/dmover/dmover.htm
31. http://www.aboutdiabetes.com/
32. http://www.blackwomenshealth.com/diabetes.htm
33. http://www.pueblo.gsa.gov/cic_text/health/noninsulin-diabetes/what.htm
34. http://www.heartfailure.org/eng_site/introheart.htm
35. http://www.lineone.net/health_encyclopaedia/ailments/pages/121/02.htm#2-
36. http://cpmcnet.columbia.edu/texts/guide/hmg16_0004.html
37. http://www.imaginis.com/heart-disease/congenital.asp
38. http://www.geneticheath.com/HD_What_Is_Heart_Disease.shtml
39. http://sln.fi.edu/biosci/healthy/attack.html
40. http://www.heartcareupdates.com/article1001.html

41. http://staff.washington.edu/bmperra/heart_help.html
42. http://www.cardiologychannel.com/heartattack/
43. http://208.133.254.45/search/display.asp?Id=417
44. http://www.healthatoz.com/atoz/heart/hearttreat.asp
45. http://www.dencats.org/heartattack/heartattack.html
46. http://cancernet.nci.nih.gov/cancertypes.html
47. http://www.maui.net/~southsky/introto.html
48. http://www.melanoma.com/whatis/index.htm
49. http://www.jas.tj/skincancer/scc.html
50. http://www.plasticsurgery.org/surgery/skincncr.htm
51. http://www.lungusa.org/diseases/lungcanc.html
52. http://www.thecancerweb.com/lung/wynk/
53. http://www.ckcc.org/treatment.htm
54. http://cancertrials.nci.nih.gov/types/leuk/blood0101.html
55. http://www.usatoday.com/life/health/cancer/lhcan000.htm
56. http://cancernet.nci.nih.gov/wyntk_pubs/brain.htm#3
57. http://www.fda.gov/fdac/features/895_brstfeed.html
58. http://www.lalecheleague.org/NB/NB6.4.90.1.html
59. http://www.epregnancy.com/info/breastfeeding/tips.htm
60. http://babiestoday.com/breastfeeding/drjack/sorenipples.htm
61. http://kidshealth.org/parent/growth/feeding/nursing_positions.html
62. http://www.nursingmoms.net/why_babies_wake.htm
63. http://www.thelaboroflove.com/forum/breastfeeding/caregiver.html
64. http://www.askdrsears.com/html/2/T028400.asp
65. http://www.healthyarkansas.com/faq/faq_breastfeeding.html
66. http://www.fourfriends.com/abrw/preparat.htm
67. http://www.twinstuff.com/breastfd4.htm
68. http://www.cota.org.au/budgethealth.htm
69. http://www.aboutseniors.com.au/HAAC.html
70. http://www.ruralhealth.gov.au/services/mps.htm
71. http://www.abs.gov.au/ausstats/abs@.nsf/94713ad445ff1425ca25682000192af2/202854896eddf986ca2569de00221c97!OpenDocument
72. http://www.uow.edu.au/arts/sts/bmartin/dissent/documents/health/access_aged.html
73. http://www.ctc.gov.au/ctc/case_studies/health_aged_care.html

Appendix 6

Health Test Set

1. http://www.cdc.gov/travel/
2. http://www.acsh.org/
3. http://www.bbc.co.uk/health/
4. http://www.intelihealth.com/IH/ihtIH/WSIHW000/408/408.html
5. http://www.viahealth.org/disease/diabetes/type1.htm
6. http://www.library.adelaide.edu.au/guide/med/menthealth/jnl.html
7. http://www.aje.oupjournals.org/
8. http://www.springerjournals.com/catalog.htm
9. http://www.heartfailure.org/eng_site/introheart.htm
10. http://www.jas.tj/skincancer/scc.html
11. http://www.diabetic.org.uk/
12. http://www.life-with-diabetes.com/
13. http://www.viahealth.org/disease/diabetes/type1.htm
14. http://www.lillydiabetes.com/Education/DiabetesTypes.cfm
15. http://www.netdoctor.co.uk/diseases/facts/diabetes.htm
16. http://www.mamashealth.com/Diabetes.asp
17. http://www.lineone.net/health_encyclopaedia/ailments/pages/121/02.htm#2-
18. http://www.diabetes.ca/about_diabetes/thefacts.html
19. http://www.healthinsite.gov.au/T.cfm?PID=1544
20. http://my.webmd.com/content/dmk/dmk_article_5963070
21. http://hlunix.hl.state.ut.us/cfhs/chronic/diabetes/New_Folder/types_of_diabete s.htm
22. http://www.heartfailure.org/eng_site/introheart.htm
23. http://www.lineone.net/health_encyclopaedia/ailments/pages/121/02.htm#2-
24. http://cpmcnet.columbia.edu/texts/guide/hmg16_0004.html
25. http://www.imaginis.com/heart-disease/congenital.asp
26. http://www.geneticheath.com/HD_What_Is_Heart_Disease.shtml
27. http://sln.fi.edu/biosci/healthy/attack.html
28. http://www.heartcareupdates.com/article1001.html
29. http://staff.washington.edu/bmperra/heart_help.html
30. http://www.cardiologychannel.com/heartattack/
31. http://208.133.254.45/search/display.asp?Id=417
32. http://www.melanoma.com/whatis/index.htm
33. http://cancernet.nci.nih.gov/cancertypes.html
34. http://www.ncbi.nlm.nih.gov/ncicgap/
35. http://www.maui.net/~southsky/introto.html
36. http://www.melanoma.com/whatis/index.htm
37. http://www.maui.net/~southsky/introto.html
38. http://www.jas.tj/skincancer/scc.html
39. http://www.plasticsurgery.org/surgery/skincncr.htm
40. http://www.lungusa.org/diseases/lungcanc.html
41. http://www.heartcareupdates.com/article1001.html
42. http://www.fda.gov/fdac/features/895_brstfeed.html
43. http://www.lalecheleague.org/NB/NB6.4.90.1.html

44. http://www.epregnancy.com/info/breastfeeding/tips.htm
45. http://www.nursingmoms.net/why_babies_wake.htm
46. http://www.thelaboroflove.com/forum/breastfeeding/caregiver.html
47. http://www.askdrsears.com/html/2/T028400.asp
48. http://www.aje.oupjournals.org/
49. http://www.cota.org.au/budgethealth.htm
50. http://www.aboutseniors.com.au/HAAC.html
51. http://www.abs.gov.au/ausstats/abs@.nsf/94713ad445ff1425ca25682000192af2/202854896eddf986ca2569de00221c97!OpenDocument
52. http://www.uow.edu.au/arts/sts/bmartin/dissent/documents/health/access_aged.html
53. http://www.melanoma.com/whatis/index.htm

BIBILIOGRAPHY

[1] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, Prabhakar Raghavan, Using Taxonomy, discriminants, and signatures for navigating in text databases. Proceedings of the 23rd VLDB Conference, Athens, Greece, 1997.

[2] P. Anick and S. Vaithyanathan, Exploiting clustering and phrases for context-based information retrieval. In SIGIR, 1997.

[3] C. Apte, F. Damerau, and S.M. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 1994. IBM Research Report RC18879.

[4] J.O. Berger. *Satistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York; Berlin, 1985, ISBN: 0-387-96098-8.

[5] L. Breiman, J.H. Friedman, R.A. Olshan, and C.J. Stone. *Classification and Regression Trees.* Wedsworth & Brooks/Cole, 1984, ISBN:0-534-98054-6.

[6] C. Chekuri, M. Goldwasser, P.Raghavan, and E. Upfal. Web search using automatic classification. Submitted for publication, 1996.

[7] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 1991.

[8] D.R. Cutting, D.R. Karger, and J.O. Pedersen. Constant interation-time scatter/gather browsing of very large document collections in SIGIR, 1993.

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

[10] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[11] C. Falaoutsos and D.W. Oard. A survey of information retrieval and filtering methods. Technical Report CS-TR-3514, University of Maryland, College Park, MD 20742, Aug 1995.

[12] W.B. Frakes and R. Baeza-Yates. *Information retrieval: Data structures and algorithms,* Prentice-Hall, 1992.

[13] K. Futunga. *An Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, New York, 1990.

[14] D. Harman. Ranking algorithms. In @.B. Frakes and R. Baeza-Yates, editors, *Information retrieval: Data structures and algorithms*, chapter 14, Prentice-Hall, 1992.

[15] D.R. Hush and B.G. Horne. Progress in supervised neural networks. IEEE *Signal Processing Magazine*, pages 8-39, January 1993.

[16] A.K. Jain, J. Mao, and K. Moiuddin. *Artificial neural networks*: A-tutorial. Computer, 29(3) 31-44, March 1996.

[17] K.S. Jones. A statistical interpretation of term specificity and its applications in retrieval. *Journal of Documentation* 28(1): 11-20, 1972.

[18] D. Koller and M. Sahami. Toward optimal feature selection In L. Saitta, editor, *International Conference on Machine Learning,* Volume 13, Morgan-kaufmann. 1996.

[19] D. Koller and M. Sahami. *Hierarchically classifying documents using very few words*. In International Conference on Machine Learning, volume 14. Morgan-Kaufmann, July 1997. To appear.

[20] P. Langley *Elements of Machine learning* Morgan Kaufman, 1996, ISBN: 1-55860-301-8.

[21] P.S. Laplace. *Philosophical Essays on Probabilities*. Springer-Verlag, New York, 1995. Translated by A.I. Dale from the 5th French edition of 1825.

[22] D. Lewis. *Evaluating text categorization*. In Proceedings of the Speech and Natural Language Workshop, pages 312-315 Morgan-Kaufmann, 1991.

[23] R.P. Lippmann. *Pattern classification using neural networks* IEEE Communications Magazine, pages 47-64, Nov 1989.

[24] B.K. Nataraja. *Machine Learning: A Theoretical Approach*, Morgan-Kaufmann, 1991.

[25] P. Papadimitriou, P. Raghavan, H. Tamaki and S. *Vempala Latent sematic indexing:* A probabilistic analysis, Submitted for publication.

[26]J. R. Quinlan. C4.5: *Programs for Machine Learning*. Morgan Kaufmann, 1993.

[27] P. Raghavan *Information Retrieval Algorithms*: A survey In Symposium on Discrete Algorithms. ACM-SIAM, 1997. Invited paper.

[28] E.S. Ristad. *A natural law of succession*. Research report CS-TR-495-95, Princeton University, July 1995.

[29] S. E. Robertson and S. Walker. *Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval*. In SIGIR, pages 232-241, 1994.

[30] G. Salton and C. Buckley. *Term weighting approaches in automatic text retrieval.* Information Processing and Management, 24(5):513-523, 1998.

[31] G. Salton and M.J. McGill *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[32] H. Schutze, D. A. Hull, and J. O. Pederson. *A comparison of classifiers and document representations for the routing problem.* In SIGIR, pages 229-237, 1995.

[33] S. Vaithyanathan. *Document classification using principal component analysis. Personal communication*, May 1997.

[34] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London. 1979. Also available online at http://www.dcs.gla.ac.uk/keith/preface.html

[35] E.M. Voorhees. *Using WordNet to disambiguate word senses for text retrieval*. In SIGIR, pages 171-180, 1993.

[36] A. Wald. *Statistical Decision functions*. Wiley, New York, 1950.

[37] S. M. Weiss and C. A. Kulikowski. *Computer Systems That Learn*. Morgan-Kaufmann, 1990.

[38] T. Y. Young and T. W. Calvert. *Classification, Estimation and Pattern Recognition*, Elsevier, 1974.

[39] G.K. Zipf *Human Behavior and the principle of Least Effort*: *An introduction to Human Ecology*. Addison-Wesley, 1949.