iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** II **Month of publication:** February 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue II, February 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Domain Based Categorisation Using Adaptive Pre-processing

Deepak Kaul[1], Nikhil Anam[2], Sourabh Gaikwad[3], Supriye Tiwari[4]
Computer Department, Savitribai Phule Pune University

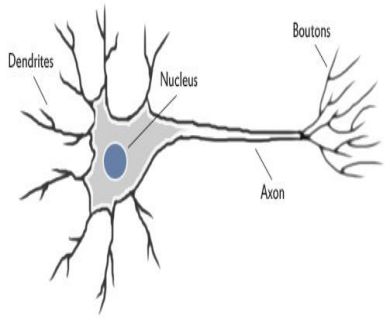
Abstract: ***As the number users accessing network for various purposes increases and simultaneously size of the Network and Internet traffic increase so, there is need for categorization web pages according to domain for easy access and also to improve the performance of system. As ANN provides Massive Parallelism, Distributed representation, Learning ability, Generalization ability, Fault tolerance. In real world applications, preprocessing plays a vital role of data mining process, as real data often comes from Different complex resources which may often noisy and redundant. So we are using an Artificial Neural Network (ANN) with adaptive pre-processing technique.***
Keywords: ***Web classification, artificial neural network, training, adaptive pre-processing, unsupervised.***

I. INTRODUCTION

In this section we analyze the two most important concepts involved for developing an automatic and adaptive classifier:

A. Artificial Neural Network

How to teach a computer ? You can either write a fixed program – or you can enable the computer to learn on its own. Living beings learn by themselves without the previous knowledge from external impressions and thus can solve problems better than any computer today. What qualities are needed to achieve such a behavior for devices like computers? There are problem categories that cannot be formulated as an algorithm. Problems that depend on many subtle factors, for example the purchase price of a real estate which our brain can (approximately) calculate. Without an algorithm a computer cannot do the same. Therefore the question to be asked is: How do we learn to explore such problems?
The answer to these questions is ANN[1].
ANN[1] imbibes ADAPTIVE[3] nature to the machine.The key to Artificial Neural Networks[1] is that their design enables them to process
information in a similar way to our own biological brains.



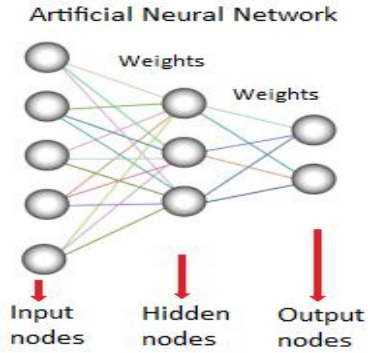
Biological neuron structure

A neuron collects inputs using a structure called dendrites, the neuron effectively sums all of these inputs from the dendrites and if the resulting value is greater than it's firing threshold, the neuron fires. When the neuron fires it sends an electrical impulse through the neuron's axon to it's buttons. These buttons can then be networked to thousands of other neurons via connections called synapses.

Components of an ANN [1]: A technical neural network consists of

1) Simple processing units,
2) The *neurons*,
3) Connections between those neurons. (directed and weighted)

Volume 4 Issue II, February 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)



The input nodes take in information, in the form which can be numerically expressed. The information is presented as activation values, where each node is given a number, the higher the number, the greater the activation. This information is then passed throughout the network. Based on the connection strengths (weights), inhibition or excitation, and transfer functions, the activation value is passed from node to node. Each of the nodes sums the activation values it receives; it then modifies the value based on its transfer function. The activation flows through the network, through hidden layers[6], until it reaches the output nodes. The output nodes then reflect the input in a meaningful way to the outside world.

A. Algorithm

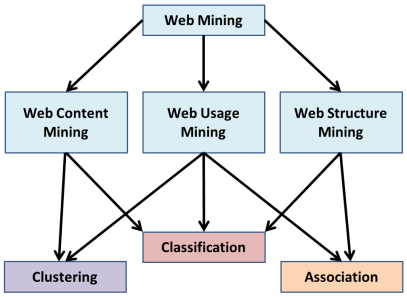
There are different types of neural networks, but they are generally classified into feed-forward and feed-back networks.

A feed-forward network is a non-recurrent network which contains inputs, outputs, and hidden layers; the signals can only travel in one direction. Input data is passed onto a layer of processing elements where it performs calculations. Each processing element makes its computation based upon a weighted sum of its inputs. The new calculated values then become the new input values that feed the next layer. This process continues until it has gone through all the layers and determines the output.Feed-forward networks are often used in data mining.

A feed-back network[6] has feed-back paths meaning they can have signals traveling in both directions using loops. All possible connections between neurons are allowed. Since loops are present in this type of network, it becomes a non-linear dynamic system which changes continuously until it reaches a state of equilibrium. Feed-back networks are often used in associative memories and optimization problems where the network looks for the best arrangement of interconnected factors.

B. Web Mining

The term Web Data Mining[4] is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest.



There are three general classes of information that can be discovered by web mining:

1) Web activity, from server logs and Web browser activity tracking.
2) Web graph, from links between pages, people and other data.

3) Web content, for the data found on Web pages and inside of documents.

While search is the biggest web miner by far, and generates the most revenue, there are many other valuable end uses for web mining results. A partial list includes:

4) Business intelligence
5) Competitive intelligence
6) Pricing analysis
7) Events
8) Product data
9) Popularity
10) Reputation

a) *Four Steps in Content Web Mining:* When extracting Web content information using web mining, there are four typical steps. When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. But web mining has additional constraints,.

11) Collect – fetch the content from the Web
12) Parse – extract usable data from formatted data (HTML, PDF, etc)
13) Analyze – tokenize, rate, classify, cluster, filter, sort, etc.
14) Produce – turn the results of analysis into something useful (report, search index, etc)

II. RELATED WORK

In this section we analyze the proposals of different research papers that we have referred.

PAPER[1] Web Page Classification based on Document Structure" This paper specifies different approaches for text or content based classification and proposes an automatic classification method based on structure of the web document.

A wide variety of techniques have been designed for text classification. Here , we will discuss the broad classes of techniques, and their uses for classification tasks. We note that these classes of techniques also generally exist for other data domains such as quantitative or categorical data. A ***Classifier***[2][1] must be ***lightweight***[2] and ***unsupervised***[2].Therefore, it is critical to design Classification methods which effectively account for these characteristics . In this chapter, we will focus on the specific changes which are applicable to the text domain. Some key methods, which are commonly used for text classification are as follows:
K-Nearest Neighbour approach [7],
Bayesian probabilistic models [4][6], 3.Inductive rule learning [5],
Support vector mechanics [1],
Neural networks [1][6]
Decision trees [1].

A. Decision Trees
Decision trees are designed with the use of a hierarchical division of the underlying data space with the use of different text features. The hierarchical division of the data space is designed in order to create class partitions which are more skewed in terms of their class distribution. For a given text instance, we determine the partition that it is most likely to belong to, and use it for the purposes of classification.

B. SVM
SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of

classification.

C. Neural Networks

Neural networks [6] are used in a wide variety of domains for the purposes of classification. In the context of text data, the main difference for neural network classifiers is to adapt these classifiers with the use of word features. We note that neural network classifiers are related to SVM[1] classifiers; indeed, they both are in the category of classifiers, which are in contrast with the generative classifiers.

D. Bayesian classifiers

In Bayesian classifiers [6](also called generative classifiers), we attempt to build a probabilistic classifier based on modeling the underlying word features in different classes. The idea is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents.

PAPER[2] CALA: An unsupervised URL-based web page classification system: This paper discusses about a URL based classifier that uses *decision trees*[] as its classifying algorithm.

Our proposal builds a number of URL patterns that represent the different classes of pages in a web site, so further pages can be classified by matching their URLs to the patterns.Its salient features are that it fulfills all of the previous requirements. There are many techniques to learn a web page classifier. They all require to download a subset of web pages, which is commonly referred to as training set, and analyze some of their features. For a learning or classification technique to be useful in the context of enterprise web information integration, it must fulfill three requirements:

1) *(R1) Lightweight crawling*[2]: To learn a classifier it is necessary to provide a training data set. Such a data set is gathered by performing a crawl of the site being analyzed. Such a crawl should be as lightweight as possible.Furthermore, web sites change frequently and it is not uncommon that these changes render the classifier obsolete. Therefore, the classifier must be learnt not once but several times and performing an extensive crawling that covers a large portion of a web site becomes unfeasible.
2) *(R2) Unsupervised*[2][3]: It is important that a person does not need to pre-classify each page in the training set individually, since this is an effort-consuming and error-prone task; neither should he or she provide any additional training information. In other words, it is important that the techniques used to learn a classifier be unsupervised or, otherwise, they shall not scale well to the Web .
3) *(R3) Classify without downloading*: Downloading a web page puts a load on the server, takes time, and consumes bandwidth. That is, it is important that a classifier relies exclusively on external features of a page in order to classify it, since it would be inefficient for practical purposes otherwise.

CALA is a proposal to learn a web page classifier that does not require a previous extensive crawling, it is unsupervised, and it does not require a page to be downloaded so that it can be classified.The system relies on two modules, namely:

1.A Crawler, which gathers a small set of hubs from a web site in order to assemble a training set;

A Pattern builder[1][2][4], which uses the previous training set to build a set of patterns, each of which represents a collection of URLs that are expected to reference web pages of the same class.

PAPER[3] Adaptive Pre-processing for Streaming Data(BASE PAPER): This paper, introduces and addresses the problem of adaptive preprocessing[3][4]. We analyze when and under what circumstances it is beneficial to handle adaptivity of preprocessing and adaptivity of the learning model[] separately. It presents three scenarios where handling adaptive preprocessing[3] separately benefits the final prediction[3] accuracy and illustrate them using computational examples. As a result of the analysis, they have constructed a prototype approach for combining adaptive preprocessing with adaptive predictor online.

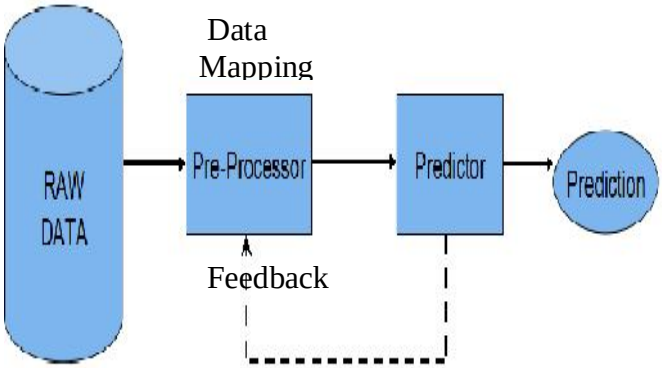
The most obvious approach to automate preprocessing in adaptive learning is to keep preprocessing tied with adaptive predictors, which can be done in two cases:

First Option: to reserve a validation set at the beginning, optimize the preprocessing parameters on that validation set, and keep the preprocessing fixed for the lifetime of the model. Only the predictor itself would adapt over time. The problem with this approach is that the system may easily fail to notice changes that happen in the raw data, and thus fail to adapt.

Second Option: to redo all preprocessing from scratch every time the predictor is retrained. This approach requires the retraining of preprocessing and a predictor to

Volume 4 Issue II, February 2016
ISSN: 2321-9653

be synchronized. That may be problematic if, for instance, preparing an accurate preprocessing requires more data than training a predictor. That may be even infeasible in cases when an incrementally adaptive predictor is used,which updates its parameters with every new instance. Last but not least, redoing the preprocessing on every data chunk may introduce unnecessary computational costs

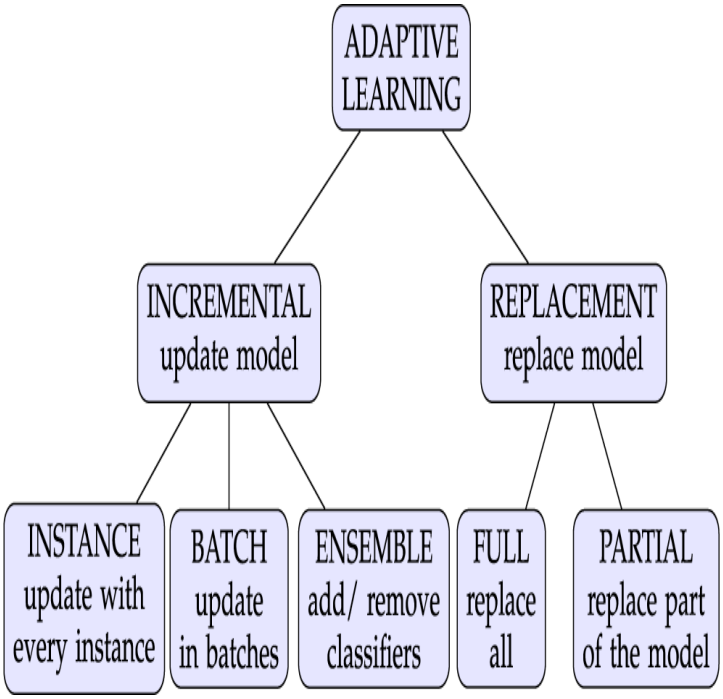


There are different types of Pre-processing

Feedback: A pre-processor[3] G may need to be learned on a training data set[2] so that it could be applied to unseen data. Learning a pre-processor can be supervised, unsupervised or independent of the data.

Operations: A preprocessor can operate in an eager way meaning that the parameters of G are fixed in the process of learning and G can be applied to new data, or lazy way, meaning that the parameters are determined from the actual incoming new data. For instance, feature extraction using PCA [8] is an eager preprocessing procedure, a rotation matrix [] can be learned on training data and fixed.

Adaptive Pre-processing Scenarios



Validation of an eager preprocessor can be organized in a direct or indirect way. Direct validation optimizes some criterion in the process of preprocessing

itself, e.g., in feature selection based on correlation with the target variable, maximization of the correlation is used as the chosen criterion. Indirect validation means that a feedback from the actual predictor is needed.

Volume 4 Issue II, February 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

III. COMPARATIVE ANALYSIS

S.NO	TITTLE	ADVANTAGES	DISADVANTAGES
1	ADAPTIVE PRE-PROCESSING FOR STREAMING DATA(BASE PAPER)	1. Decoupled Adaptive Preprocessing and Adaptive Learning. 2.Increased accuracy (above 80%) 3.Automates the machine Learning process.	1.Implementation complexity 2.Association rule not universal 3.Varying learning rates of Preprocessors & Predictors
2	WEB PAGE CLASSIFICATION BASED ON DOCUMENT STRUCTURE	1.Avoids irrelevant Information 2.Query string constructed using hierarchically organized info 3.Diverse classification range	1.Only image and structural info is fetched 2.Requires extensive downloading 3.Less effective machine learning Algorithm
3.	CALA: AN UNSUPERVISED URL-BASED WEB PAGE CLASSIfiCATION SYSTEM	1.Highly suitable for enterprise web info Integration 2.no need 3.Lightweight 4.Zero intervention in site Operation	1.Preprocessed data not completely relevant 2.Initial web clustering done manually 3.Accuracy of p-estimator debatable. 4.SVM -high algorithmic complexity and memory requirement.

IV. CONCLUSION

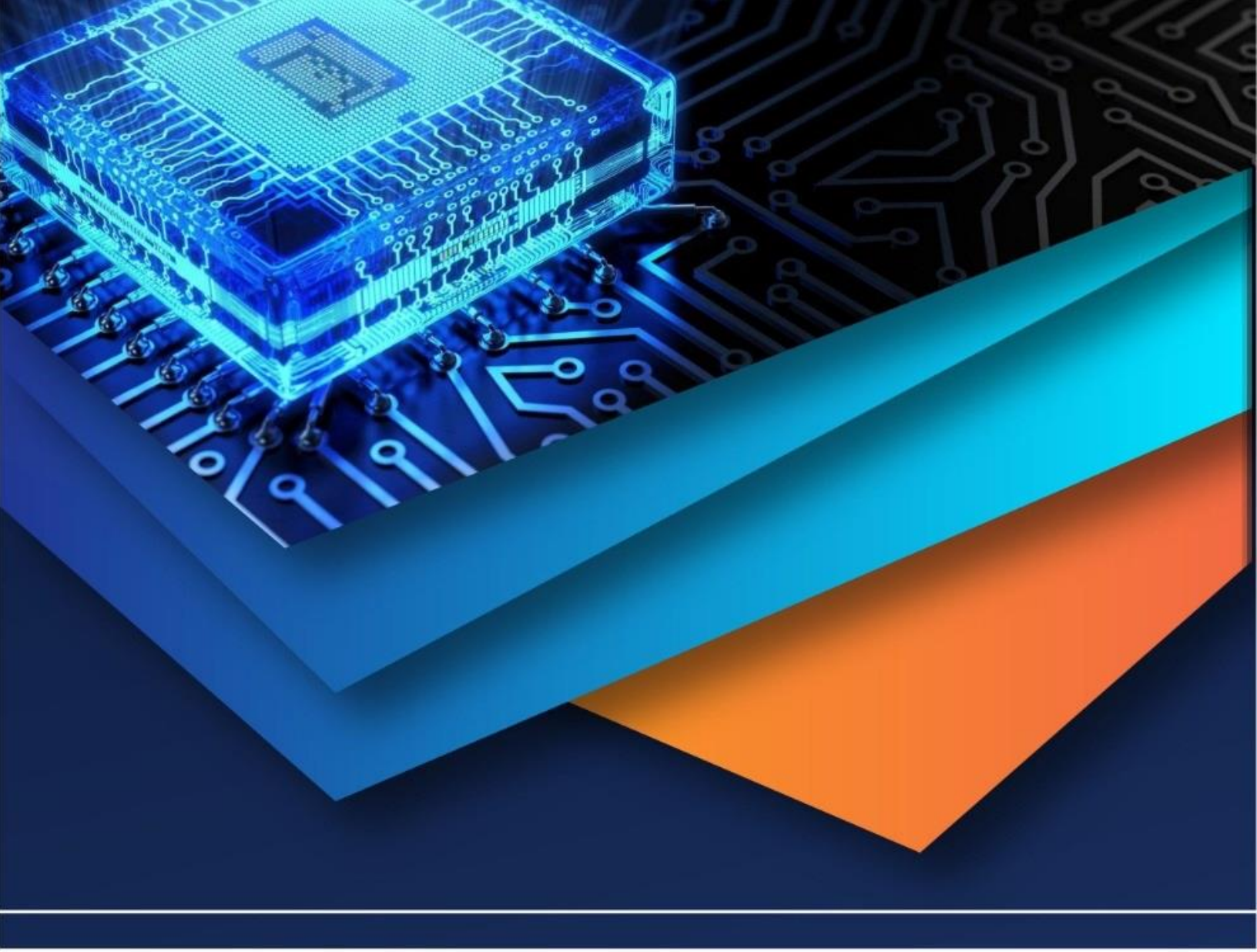
Any efficient classifier needs to be unsupervised and its crawler must be lightweight.Ann(artificial neural network) is the best machine learning algorithm for automating and providing decisive abilities to a classifier.A neural network based classification approach could be employed to automate the training process. Adding a few more features based on heuristics, (e.g. the classification of a page as a home page by detecting a face at the top) would increase the classification accuracy. An ANN[1] based classifier that uses Adaptive Preprocessing[3] has a no of advantages such as least user intervention,low time consumption and self developing knowledge to categorize a Domain[2].

V. ACKNOWLEDGMENT

We would like to express my gratitude and appreciation to all those who gave us the possibility to complete this report. A special thanks
to our project coordinator, Mrs. Urmila Biradar,and Project guide Mrs.Arti Deshpande whose help, stimulating suggestions and encouragement,helped me to coordinate my project and especially in writing this report.

REFERENCES

[1] Arul Prakash Asirvatham, Kranthi Kumar. Ravi "Web Page Classification based on Document Structure", 2012,http://www.cs.utah.edu/~arul/papers/web_page_classification.pdf
[2] I.Hernandez.."CALA An unsupervised URL-based web page classification system",2013, http://www.elsevier.com/locate/knosys
[3] Indre Zliobaite Bogdan Gabry "Adaptive Preprocessing for Streaming Data",2014,http://dl.acm.org/citation.cfm
[4] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda,"New Ensemble Methods for Evolving Data Streams," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining(KDD '09), pp. 139-148, 2009.
[5] Susan Dumais, Hao Chen, Hierarchial Classification of web content
[6] John.M.Pierre, Practical Issues for Automated Categorization of Web Pages,September 2000
[7] Oh-Woog Kwon, Jong-Hyoek Lee, Web page classification based on k-Nearest Neighbor approach



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY