Automatic Derivation of On-line Document Ontologies

Mechanisms for Enterprise Integration: From Objects to Ontologies - MERIT 2001
Budapest, Hungary
Jun 19th, 2001

J.Rafael G.Pulido[1], Dave Elliman
[jrp|dge]@cs.nott.ac.uk
Nottingham University
Computer Science and IT School
United Kingdom

Sponsored by
PROMEP (SEP-CONACyT)
Colima University
Mexico

[1] Corresponding Author

CONTENTS

- Introduction
- Related Work
- Our Research
- Conclusions

INTRODUCTION

Motivation

- The WWW is growing rapidly
- Vast amounts of text are now available in machine-readable form and can be processed electronically
- Web users want effortlessly to locate the right piece of information

but, it is sometimes virtually impossible to locate useful items

Several authors have studied these problems:

- how to extract **knowledge** from documents
  [Guarino, 1998]
  [Borst et al., 1996]
  [Gangemi et al., 1998]

- how to organize it
  [Salton et al., 1996]
  [Sanz et al., 1998]
  [Ackerman and Fielding, 1995]

- how to deliver it to users
  [Kohonen et al., 1999]
  [Merkl, 1999]
  [Salton, 1968]

however, most of the literature treats the problem in an isolated way

Approach

By using a combination of two relatively recent techniques:

- Ontologies
- Self-Organizing Maps

our method aims to extract knowledge from digital sources, and to create browsable and reusable collections of it

This research may also provide an alternative to:

- solve the problem of classifying electronic information
- the way in which electronic archives can be explored
- how knowledge can be extracted from them and shared with external software agents

Pattern Recognition (PR)

Humans perform everyday tasks in such a way that they may seem simplistic and uncomplicated

PR is the study of how machines can observe their environment, distinguish patterns from their background, and make decisions about the categories of the patterns [Jain, 2000]

For instance, people can often be able to identify smells, voices, or even faces, in spite of some variations like rotation or illumination, received through the senses from the surrounding environment

It would seem an easy task to create machines with those brain-like skills

However, as [Wasserman, 1989] says, nothing could be further from the truth

Nowadays, differences and similarities, obvious for humans, still confuse the most sophisticated PR Systems

The exact manner in which this complex system accomplishes such tasks is little understood, however much of the physiological structure has been mapped, and certain functional areas are gradually yielding to determine research.

Artificial Neural Networks is one of many approaches to perform Pattern Recognition

Examples of patters[2] are fingerprint images, handwritten words, human faces, speech signals, **text documents**, and the like

[2] A pattern can be described as an entity that could be given a name

Artificial Intelligence (AI)

AI[3] is the study of the computations that make possible to perceive, reason, and act [Winston, 1992]

It has the same objectives that Neural Networks[4] have, and both share the effort that researches are doing to build more useful and intelligent machines

[3] models theories rooted in Psychology

[4] models the biology putting more effort to how the brain functions

Methods

The most popular ones in dealing with PR tasks are:

Template Matching one of the simplest and earliest approaches; it determines the similarity between two entities of the same type [5]

Statistical Classi$f$ication each pattern is represented in terms of a number of features, which is viewed as a point in a multi-dimensional space

Syntactic Matching provides a description of how the given pattern is constructed from *primitives* [6]

A more complex pattern can be outlined in terms of the interrelations between these primitives

[5] a prototype of the pattern to be recognized must be available

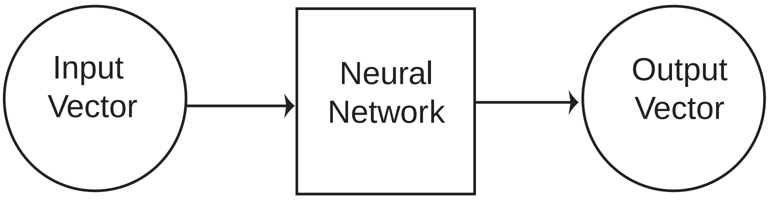
[6] the simplest subpatterns to be recognized

Artificial Neural Networks (NN)

NN can be described as:

A parallel, distributed information processing structure consisting of elements interconnected together with unidirectional signal channels called connections, where each element has a single output connection, which branches into as many collateral connections as desired, and where the output can be any mathematical type desired [Simpson, 1990]

Single Layer Neural Network



They are also known as:

- Neuro-Computing Systems
- Parallel Distributed Processing Models
- Adaptive Systems

They are biologically inspired, and try to imitate the way in which the brain’s cells work

Although small in degree, this emulation of the brain has produced some impressive results, because they have the ability to:

- learn from experience
- generalize from their knowledge
- effectuate abstractions
- make errors

all these more characteristic of human thought that computers

Training

In order to learn, Neural Networks have to be trained

In traditional programming, the input-output relationship is established beforehand by the analyst

In contrast, Neural Networks do not require instructions, rules or processing requirements about how to process the data

They determine, in most of the cases, relationships by looking at examples of input-output pairs

There are two types of training:

Supervised training involves a teacher, and requires input vectors to be paired with $target$[7] vectors

Unsupervised training [8] this training does not need a teacher, and extracts features from the inputs themselves

After this stage of training, either supervised or unsupervised, the system is ready to use

[7] desired outputs

[8] no target is especified a priori

Ontologies

Ontologies is one of the techniques that we are using in trying to solve the problem

Its importance has been recognized in disciplines as diverse as knowledge engineering, knowledge representation, qualitative modelling, language engineering, database design, information integration, object-oriented analysis, information retrieval and extraction, knowledge management and organization, and agent-based systems design [Guarino, 98]

What are they?

- A branch of metaphysics concerned with the nature and relations of being, and about the nature of being or the kinds of existents [9]

- [Borst et al., 1996]:

  Primary Ontologies represent distinguished parts of the world, modelling general views on a particular domain (concepts and relations relevant to a certain domain)

  Secondary Ontologies introduce additional distinctio or typologies that can be applied to objects of primary Ontologies

[9] Merriam-Webster Dictionary Online at http://www.m-w.com

[Guarino, 1998] divides it as follows:

Theory of parts consists of the parts of a given entity, its properties and the different kind of parts

Theory of wholes which parts of a whole are interconnected, and the properties of such connections

Theory of identity describes how an entity can change, while it maintains its properties, and when an entity loses its individuality

Theory of dependence studies the various forms of existential dependence, which involves particular individuals that belong to different classes

The main idea behind Ontologies, is that people intrinsically use them for classifying ideas and concepts in their brains

The same idea supports the effort that different groups of researchers are doing all over the world

RELATED WORK

Related Work

Reseach from a number of areas has been used to document this study, namely:

- Text Analysis, and Natural Language Processing
- Neural Networks, and Artificial Intelligence
- Web Searching, Digital Libraries
- Ontologies, Knowledge-Based Systems
- Self-Organizing Maps, and Pattern Recognition
- Object-Oriented Technology

Gatherers

In [Sanz et al., 1998] a distributed architecture for the extraction of meta-data from WWW documents is proposed which is particularly suited for repositories of historical publications

This information extraction system is based on semi-structured data analysis

Gatherers have been designed as a combination of a parser, based on a context-free grammar, and a web robot, which navigates the links contained in the basic document type to infer the document structure of the entire site

Meta-data objects can be classified and organized, then interchanged with other web agents

Simple HTML Ontology Extension

In [Gangemi et al., 1998], the use of **SHOE** in a real world internet application is described

This approach allows authors to add semantic content to web pages, relating the context to common ontologies that provide contextual information about the domain

Most web pages with SHOE $annotations$ tend to have tags that categorize concepts

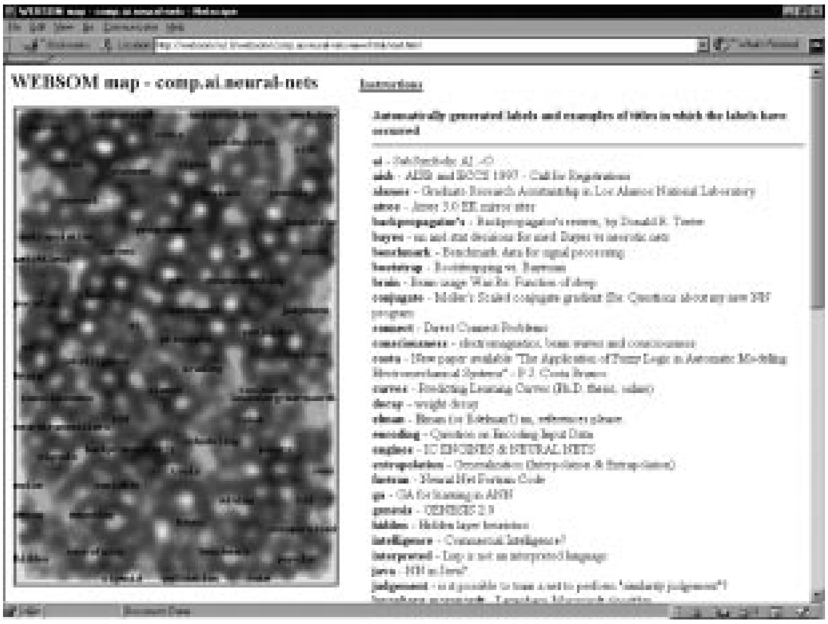
therefore, **there is no need for complex inference rules to perform automatic categorization**

WEBSOM2

It is a document organization, searching and browsing system, which has has been used to a set of about 7 million of electronic patent abstracts

It presents a document map as a series of HTML pages enabling exploration

Best-matching points are marked with a symbol, which can be used as starting points for browsing [10]


WEBSOM map - comp.ai.neural-nets - Netscape
File Edit View Go Communicator Help
Bookmarks
Location
What's Related

WEBSOM map - comp.ai.neural-nets

Instructions

Automatically generated labels and examples of titles in which the labels have occurred

ai - Sub-Symbolic AI :-O
aisb - AISB and ECCS 1997 - Call for Registrations
alamos - Graduate Research Assistantship in Los Alamos National Laboratory
atree - Atree 3.0 EK mirror sites
backpropagator's - Backpropagator's review, by Donald R. Tveter
bayes - nn and stat decisions for med: Bayes vs neurotic nets
benchmark - Benchmark data for signal processing
bootstrap - Bootstrapping vs. Bayesian
brain - Brain usage Was Re: Function of sleep
conjugate - Moller's Scaled conjugate gradient (Re: Questions about my new NN program)
connect - Direct Connect Problems
consciousness - electromagnetics, brain waves and consciousness
costa - New paper available "The Application of Fuzzy Logic in Automatic Modelling Electromechanical Systems" - F.J. Costa Branco
curves - Predicting Learning Curves (Ph.D. thesis, online)
decay - weight decay
elman - Elman (or Edelman?) nn, references please
encoding - Question on Encoding Input Data
engines - IC ENGINES & NEURAL NETS
extrapolation - Generalization (Interpolation & Extrapolation)
fortran - Neural Net Fortran Code
ga - GA for training in ANN
genesis - GENESIS 2.0
hidden - Hidden layer heuristics
intelligence - Commercial Intelligence?
interpreted - Lisp is not an interpreted language
java - NN in Java?
judgement - is it possible to train a net to perform "similarity judgements"?

Document: Done

[Kohonen et al., 1999]

[10] http://websom.hut.ƒi/websom

The Smart System

Some facilities that this system included:

- Stemming
- Synonym dictionary
- **Hierarchical arrangement**
- Statistical association
- Syntactic analysis
- Statistical phrase

[Salton, 1968]

Statistical Associations

$$\underbrace{\overbrace{aircraft = \left[\begin{matrix} airframe \\ fuselage \\ propeller \\ supersonic \\ acceleration \end{matrix}\right.}^{1st\ order\ associations(aircraft-airframe)} \qquad = airplane}_{2nd\ order\ associations(aircraft-airplane)}$$

Higher-Order Associations

A term-term similarity matrix can be described as

$$\mathbf{C}^T \cdot \mathbf{C} \tag{1}$$

while, a doc-doc similarity matrix as

$$\mathbf{C} \cdot \mathbf{C}^T \tag{2}$$

These matrices specify $first-order$ associations between objects

Higher-order associations can be produced by generating $(\mathbf{C}^T \cdot \mathbf{C})^n$ and $(\mathbf{C} \cdot \mathbf{C}^T)^n$
[Salton, 1968]

A Similarity Measure

In order to be able to make a decision concerning the closeness between two documents, it is necessary to compute a *ratio* between them

$$c_{ij} = \frac{\sum\limits_{k} min(d_k^i, d_k^j)}{\sum\limits_{k} d_k^i} \tag{3}$$

$$c_{ij} = \begin{cases} 1 & , \quad d^i = d^j \\ 0 & , \quad d^i \neq d^j \end{cases} \tag{4}$$

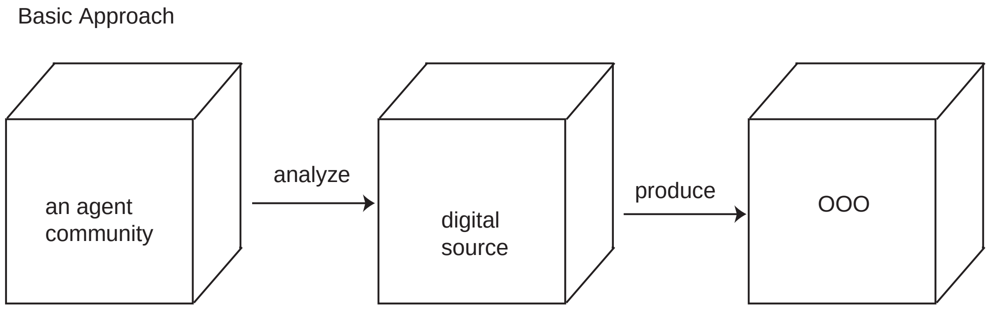
where d^i, d^j are k-dimensional property vectors representing terms w_i, w_j

OUR RESEARCH

Our Research

An ontology is a form of knowledge representation that provides a common vocabulary of concepts and relationships which may be used to inform a viewer, a search engine, or other software entities

We aim to create a browsable collection of knowledge [11] for a particular digital source



[11] Object Oriented Ontology - OOO

This browsing tool can be outlined as a number of sets [Salton, 1968]:

set of objects (entities, concepts)

set of functions (for example: is-a)

set of relations (between objects)

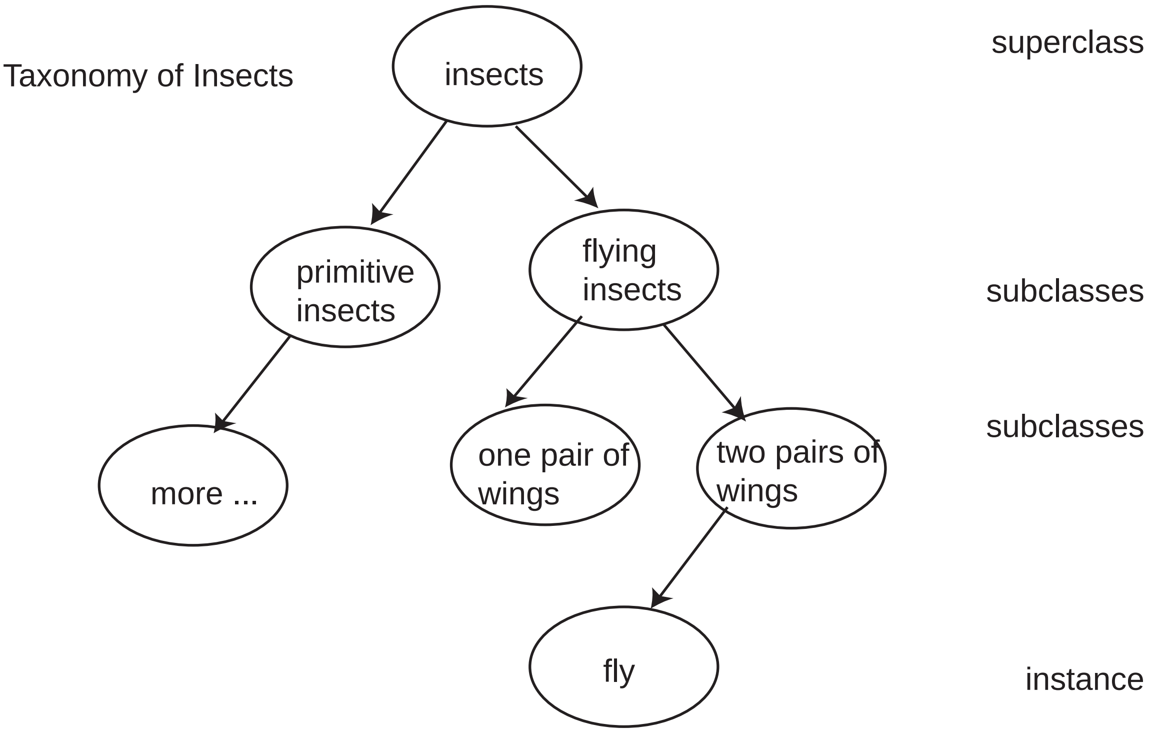
set of semantic rules [12]

[12] no that rigid though

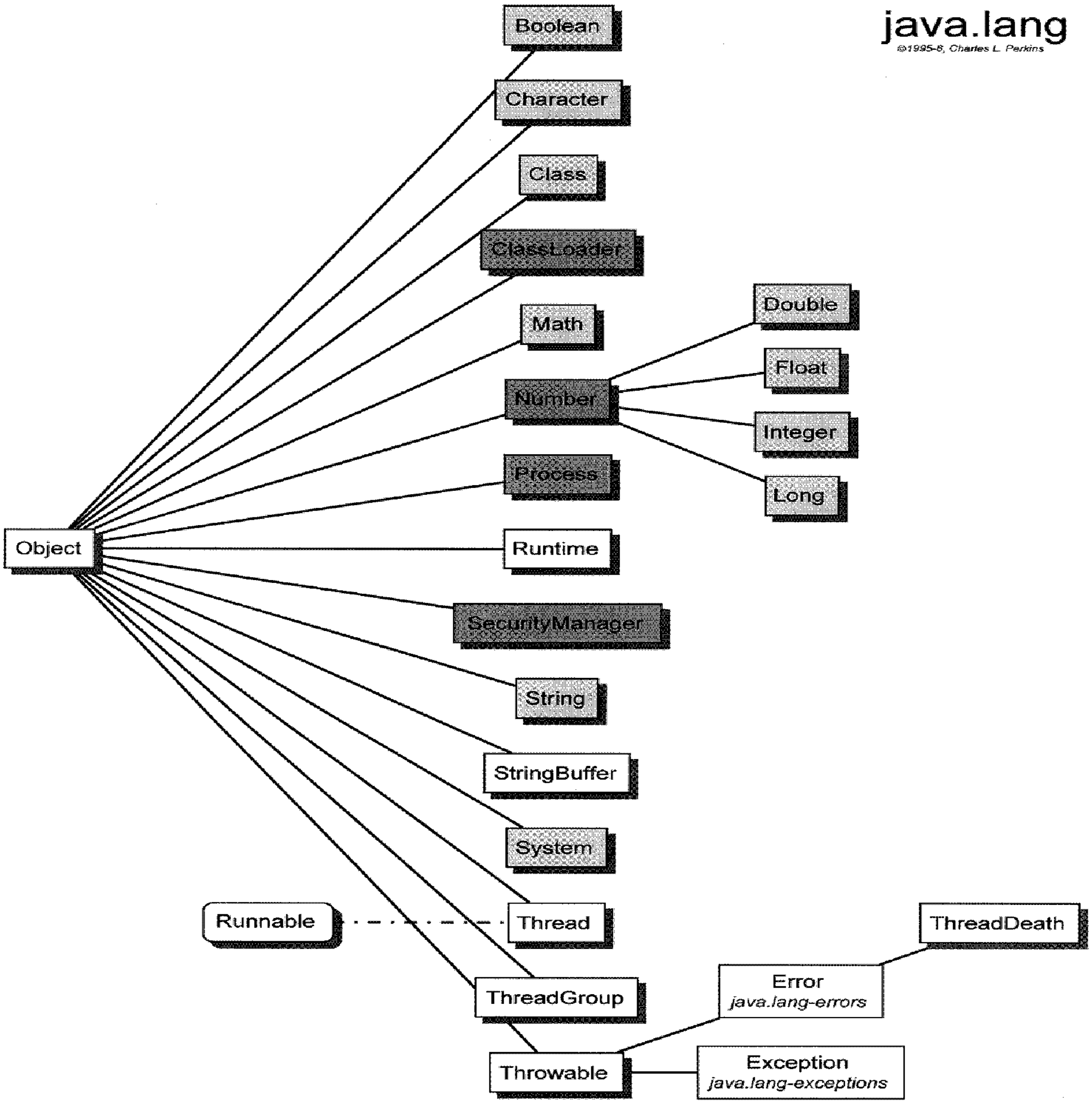
Hierarchy Concept

Entomologists classify insect using a system called *taxonomy*

Take the case of Insects, they are ordered by families and classes



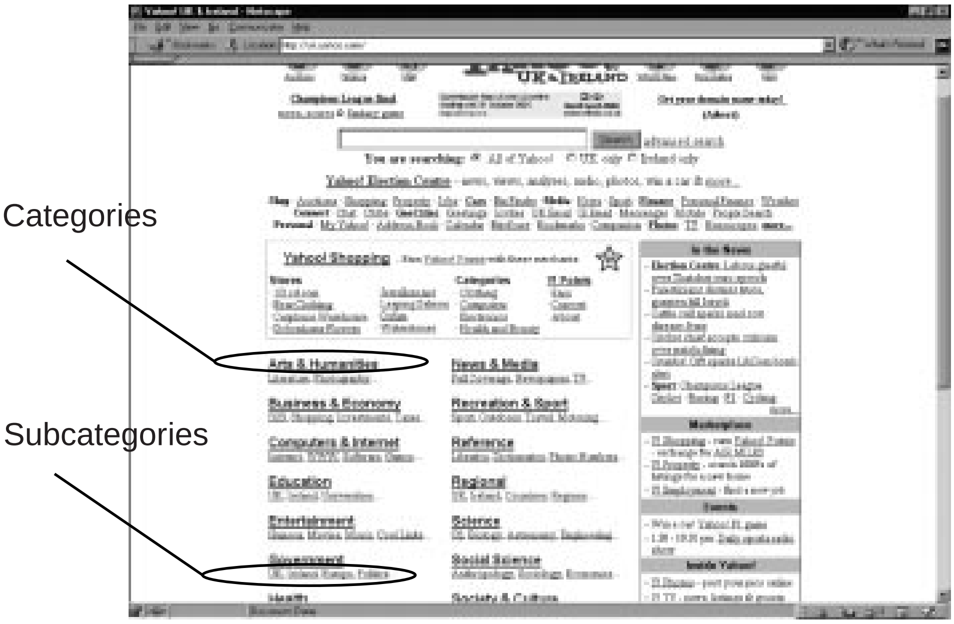
A Java Taxonomy



[Lemay and Morrison, 1996]

A more familiar Taxonomy

Yahoo[13] presents categories and subcategories



Categories

Subcategories

Arts & Humanities
Literature, Photography...

News & Media
Full Coverage, Newspapers, TV...

Business & Economy
B2B, Shopping, Investments, Taxes...

Recreation & Sport
Sport, Outdoors, Travel, Motoring...

Computers & Internet
Internet, WWW, Software, Games...

Reference
Libraries, Dictionaries, Phone Numbers...

Education
UK, Ireland, Universities...

Regional
UK, Ireland, Countries, Regions...

Entertainment
Games, Movies, Music, Cool Links...

Science
UK, Biology, Astronomy, Engineering...

Government
UK, Ireland, Europe, Politics...

Social Science
Anthropology, Sociology, Economics...

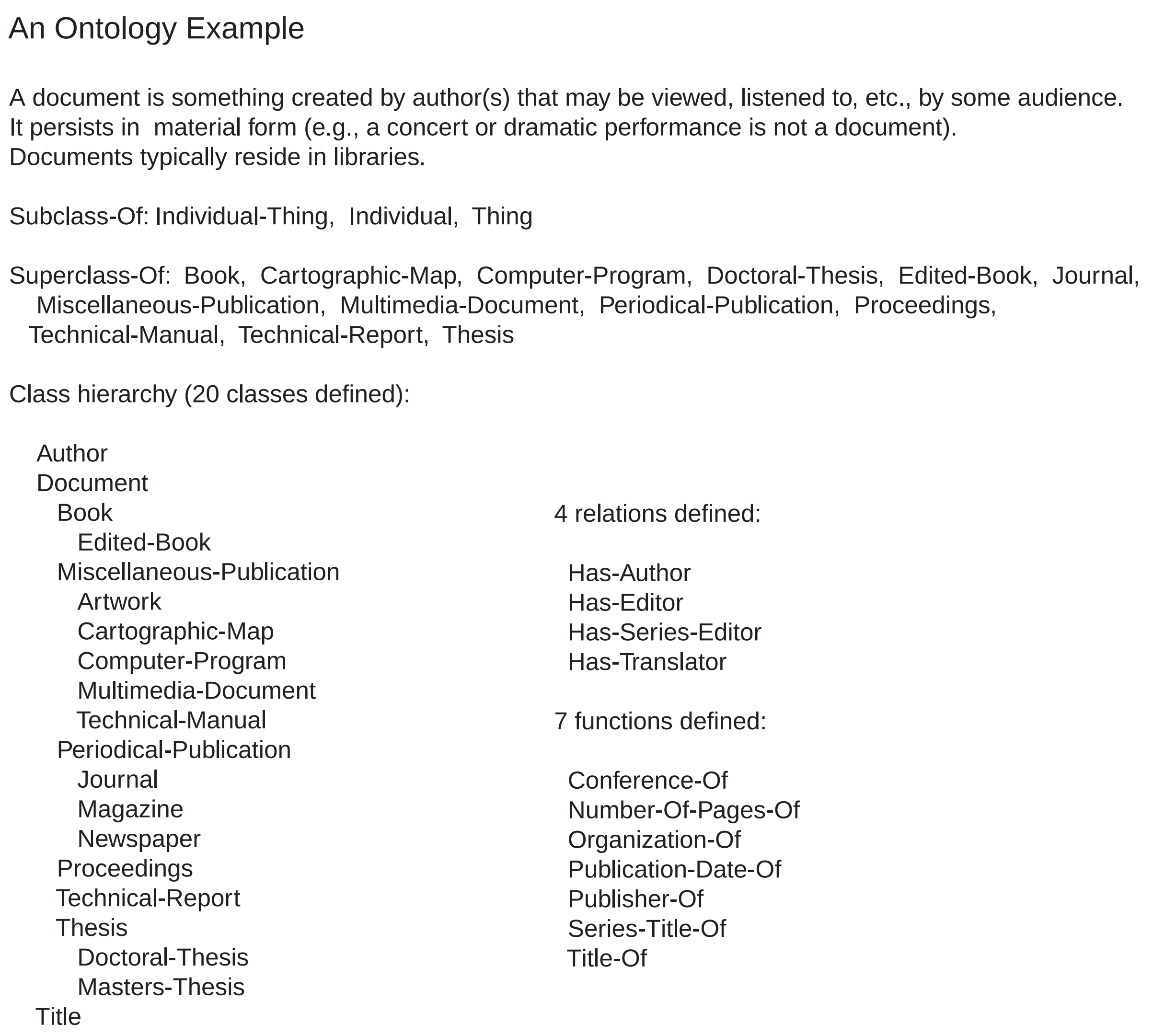
Health

Society & Culture

[13] http://uk.dir.yahoo.com

Piece of ’ontology’

Document ontology described at Stanford KSL Network Services [14].



An Ontology Example

A document is something created by author(s) that may be viewed, listened to, etc., by some audience. It persists in material form (e.g., a concert or dramatic performance is not a document). Documents typically reside in libraries.

Subclass-Of: Individual-Thing, Individual, Thing

Superclass-Of: Book, Cartographic-Map, Computer-Program, Doctoral-Thesis, Edited-Book, Journal, Miscellaneous-Publication, Multimedia-Document, Periodical-Publication, Proceedings, Technical-Manual, Technical-Report, Thesis

Class hierarchy (20 classes defined):

- Author
- Document
  - Book
    - Edited-Book
  - Miscellaneous-Publication
    - Artwork
    - Cartographic-Map
    - Computer-Program
    - Multimedia-Document
    - Technical-Manual
  - Periodical-Publication
    - Journal
    - Magazine
    - Newspaper
  - Proceedings
  - Technical-Report
  - Thesis
    - Doctoral-Thesis
    - Masters-Thesis
- Title

4 relations defined:

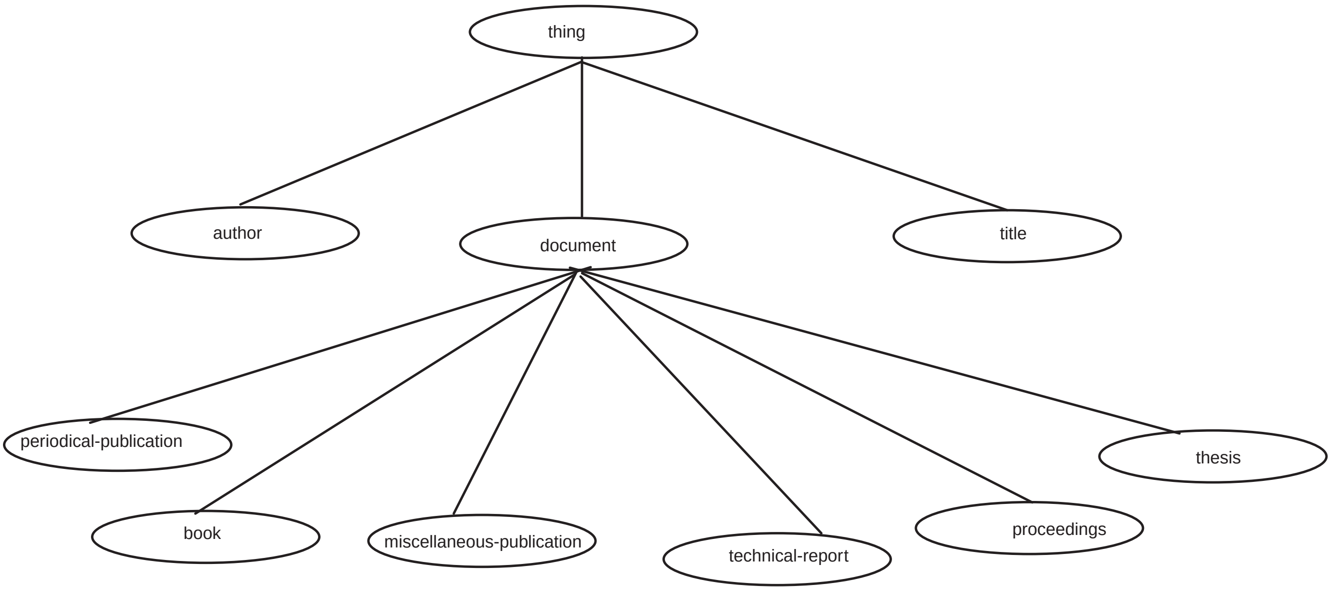
- Has-Author
- Has-Editor
- Has-Series-Editor
- Has-Translator

7 functions defined:

- Conference-Of
- Number-Of-Pages-Of
- Organization-Of
- Publication-Date-Of
- Publisher-Of
- Series-Title-Of
- Title-Of

[14] http://www-ksl-svc.stanford.edu:5915

Taxonomy Derived

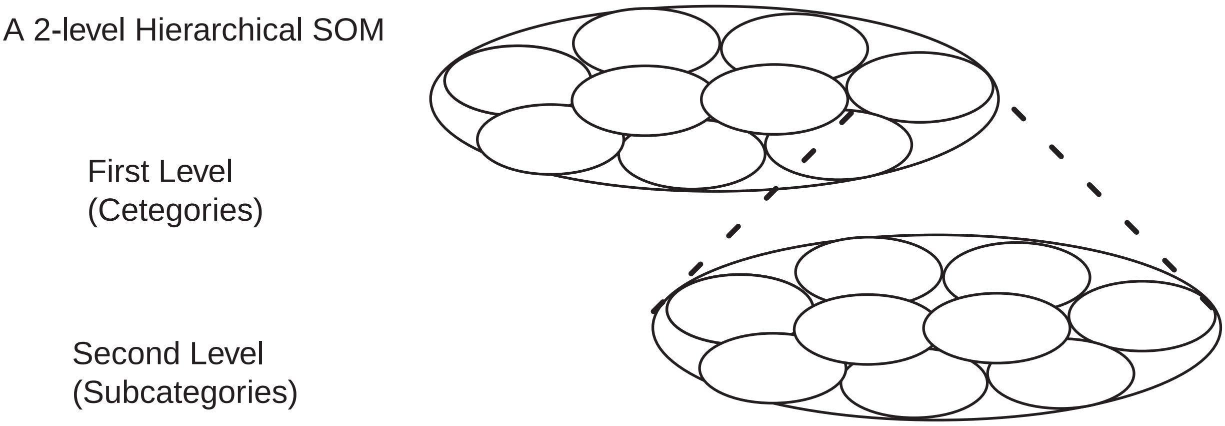


Hierarchical Feature Maps

Consist of a number of individual self-organizing maps, and represent the contents of a document archive in form of a taxonomy [Merkl, 1999]

The distinguished feature of this model is that it provides a hierarchical view of the underlying data collection in form of an atlas where, starting from a map representing the complete data collection, different levels are shown at $finer$ levels of granularity

The complete archive is represented by means of a small overview map



A 2-level Hierarchical SOM

First Level
(Cetegories)

Second Level
(Subcategories)

The Programming Language Java

Java has a neutral architecture and is a portable language, and was created to meet the requirements of networked, object oriented, multi-threaded systems from the start, not as an afterthought

Some important features of Java are:

- Object Oriented Programming Language
- Multi-platform Computing
- Client/Server Computing
- **Parallel Computing**

Our attention, however, is focused on these ones:

Object Oriented Programming Language Object Oriented Technology is widely accepted among developers because it helps model real world entities in terms of software components

Parallel Computing Most of the software that we usually write is single-threaded

Java, on the other hand, supports multi-thread programming, which means that more that one concurrent sequence instruction can be running simultaneously

Text Categorization

These are some text categorization and dimensionality reduction tools:

- Document Frequency (DF)
- Category Frequency (CF)
- Term Frequency times Inverse Document Frequency (tf x idf)
- Principal Component Analysis (PCA)
- Self-Organizing Maps (SOM)

[Lam and Lee, 1999]

Vector Space Basic Model

Documents can be regarded as high dimensional vector spaces $d_i = \{w_{i1}, w_{i2}, \cdots, w_{in}\}$

d	w_1	w_1	$\cdots$	w_n
d_1	7	5	$\cdots$	1
d_2	1	3	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_n	2	11	$\cdots$	10

$\{w_1, w_1, \cdots, w_n\}$ are *terms*, and $\{d_1, d_2, \cdots, d_n\}$ *documents*

$$\mathbf{n}_j = \sum_k d_{jk}\mathbf{e}_k \tag{5}$$

where

$\mathbf{e}_k$: unit vector

d_{jk} : frequency of occurrence of w_j in d_k

Random Mapping

A new rapid dimensionality reduction technique

$$\mathbf{x}_j = \sum_k d_{jk} \mathbf{r}_k \qquad (6)$$

where
$\mathbf{r}_k$: lower-dimensional random vector
d_{jk} : frequency of occurrence of w_j in d_k

The above equation can be expressed as a simple matrix multiplication

$$\mathbf{x}_j = \mathbf{R}\mathbf{n}_j \qquad (7)$$

$\mathbf{n}_j \in \mathcal{R}^N$
$\mathbf{x}_j \in \mathcal{R}^n$

$n \ll N$
[Kaski, 1998]

Binary Vector Space

By specifying a *cut-off* value, vector spaces can be transformed into binary ones

d	w_1	w_1	$\cdots$	w_n
d_1	1	0	$\cdots$	1
d_2	1	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_n	0	1	$\cdots$	1

Two documents, or terms, are related if and only if the similarity factor is at least equal to a given threshold

Term Frequency times Inverse Collection Frequency (tf x icf)

$$idf_i = \log \frac{N}{n} \tag{8}$$

N : size of the document collection
n : number of documents in the collection with term w_k

$$w_{ik} = \frac{tf_{ik} \cdot \log(\frac{N}{n_k})}{\sqrt{\sum_{j=1}^{t} (tf_{ij})^2 \cdot (\log \frac{N}{n_j})^2}} \tag{9}$$

w_{ik} : weight of term w_k in document d_i

tf_{ik} : frequency of occurrence of term w_k in d_i

N : size of the document collection

n_k : number of texts in the collection with term w_k

$\log \frac{N}{n_k}$: inverse collection frequency factor

[Salton et al., 1996]

Principal Component Analysis

While td x idf is used for textual data, PCA can be applied to a wide variety of data

From a feature space **x**, a set of orthogonal unit vectors are found **u**, to form a set of uncorrelated projections **a**, called principal components

$$a_i = \mathbf{x}^T \mathbf{u}_i \tag{10}$$

A reduced feature space is formed by taking the prime p components with the largest variance

Self-Organizing Maps

SOMs are a new, effective software tool for the visualization of high-dimensional data

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)] \tag{11}$$

$$h_{ci}(t) = \alpha(t)\exp(-\frac{||\mathbf{r}_c - \mathbf{r}_i||^2}{2\sigma(t)}) \tag{12}$$

where
$\mathbf{m}_i \in \mathcal{R}^n$ (the winner)
$h_{ci}(t)$ (neighbourhood function)
$\mathbf{x}, \mathbf{m}_i \in \mathcal{R}^n$
$\mathbf{r}_c, \mathbf{r}_i \in \mathcal{R}^2$
$0 \leq \alpha \leq 1$
$\alpha, \sigma \to 0$
$c = \min ||\mathbf{x} - \mathbf{m}_c||$
[Kohonen, 1998]

Constructing Ontologies

A set of documents are processed to build an ontology in terms of a user $specified$ query

After html-untagging, a $stoplist$ of common words that carry little information is used to prune these words from documents, then stemming

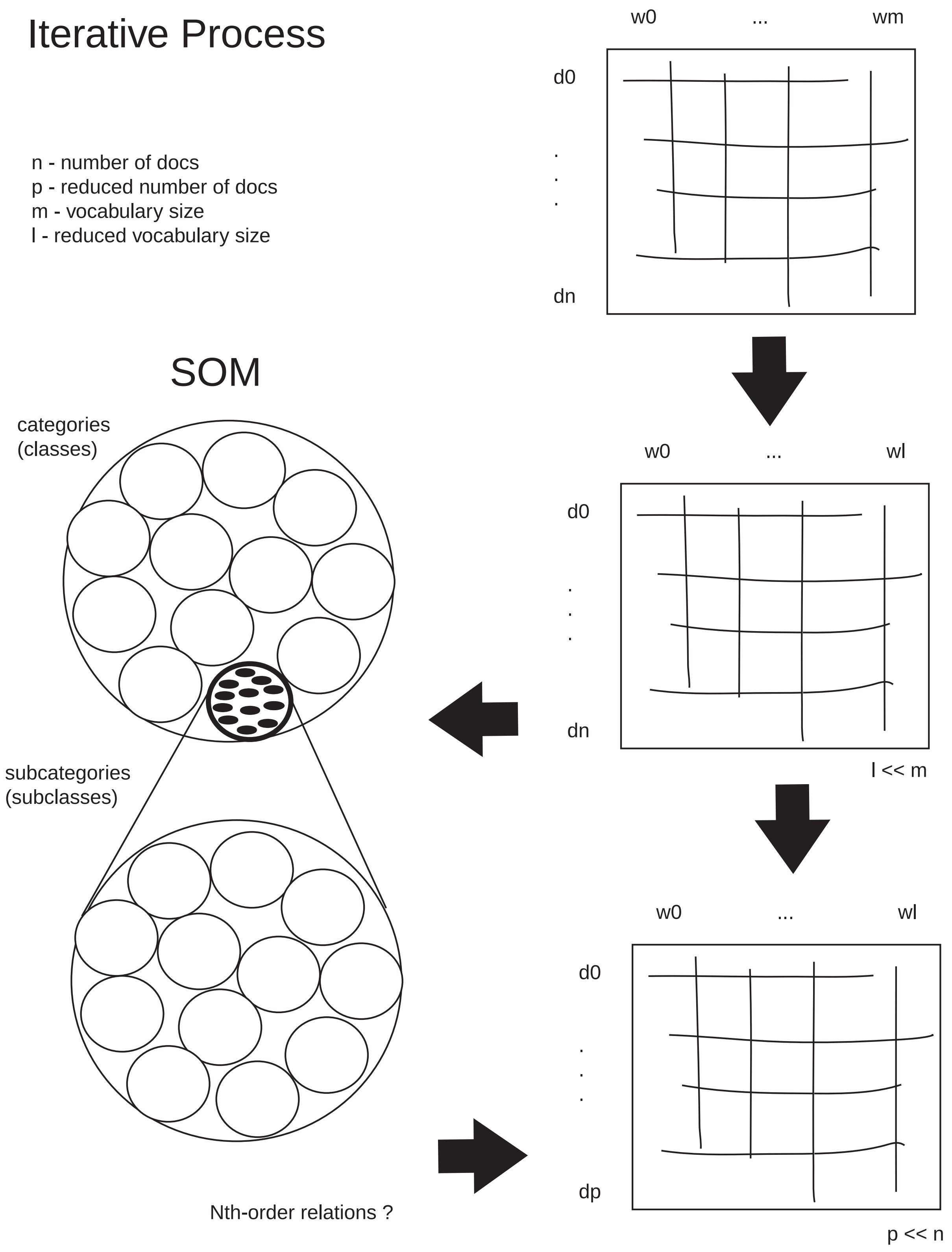
An initial ontology was constructed by counting the occurrence per thousand words of the hypernyms of the search keywords or their synonyms

A document vector is then formed from the remaining words, applying text $encoding$ and a principal component analysis is performed to identify the key words that are most powerful in separating individual documents

The foremost terms (50) are taken to form eigenvectors, and these are combined with the user information to train a SOM, which produced an alternative structured view of the information which was suitable for display to the user

Iterative Process

n - number of docs
p - reduced number of docs
m - vocabulary size
l - reduced vocabulary size



CONCLUSIONS

Conclusions

There is a growing need for methods of systematic explorative information retrieval, where the exact keywords which could guide to relevant and interesting information, may not be known in advance

Clearly, Ontologies are being used as tools for Knowledge Engineers in the task of representing knowledge

Ontologies can be used to automate processes, help people understand others' ideas, create knowledge to be used by either robots, or software agents, and elicit knowledge from electronic documents

Ontologies can be used to give a sense of order to unstructured digital sources such as the web

What’s Next

Further research in this area is needed in order to produce new and improved techniques to better help computer users locate the information they need

We are investigating the use of hierarchical SOMs and the association of further keywords from HTML tags and from the hyperlink structure of the web site itself

Automatic Derivation of On-line Document Ontologies

Mechanisms for Enterprise Integration: From Objects to Ontologies - MERIT 2001
Budapest, Hungary
Jun 19th, 2001

J.Rafael G.Pulido[15], Dave Elliman
[jrp|dge]@cs.nott.ac.uk
Nottingham University
Computer Science and IT School
United Kingdom

Sponsored by
PROMEP (SEP-CONACyT)
Colima University
Mexico

[15] Corresponding Author

References

[Ackerman and Fielding, 1995] Ackerman, M. and Fielding, R. (1995). Collection maintenance in the digital library. In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries*, pages 39 48.

[Borst et al., 1996] Borst, P., Benjamins, J., Wielinga, B., and Akkermans, H. (1996). An application of ontology construction. In *Workshop on Ontological Engineering, ECAI'96*, pages 5 16.

[Gangemi et al., 1998] Gangemi, A., Pissanelli, D., and Steve, G. (1998). Ontology integration: Experiences with medical terminologies. In *Formal Ontology in Information Systems*, pages 163 178. IOS.

[Guarino, 1998] Guarino, N. (1998). Some ontological principles for designing upper level lexical. In *the First International Conference on Lexical Resources and Evaluation*.

[Guarino, 98] Guarino, N. (98). Formal ontology and information systems. In *Formal Ontology in Information Systems*, pages 3 15. IOS Press.

[Jain, 2000] Jain, A. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence Journal*, 22(1):4 37.

[Kaski, 1998] Kaski, S. (1998). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, volume 1, pages 413 418. IEEE.

[Kohonen, 1998] Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21.

[Kohonen et al., 1999] Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., and Saarela, H. (1999). Self organization of a massive text document collection. In *Kohonen Maps*, pages 171 182. Elsevier Science.

[Lam and Lee, 1999] Lam, S. and Lee, D. (1999). Feature reduction for neural network based text categorization. In *6TH Conference On Database Systems For Advanced Applications*, pages 195 202. IEEE.

[Lemay and Morrison, 1996] Lemay, L. and Morrison, M. (1996). *Teach Yourself Java in 21 Days, Professional Reference Ed.* SAMS.

[Merkl, 1999] Merkl, D. (1999). Document classification with self-organizing map. In *Kohonen Maps*, pages 183 192. Elsevier Science.

[Salton, 1968] Salton, G. (1968). *Automatic Information Organization and Retrieval*. McGraw-Hill.

[Salton et al., 1996] Salton, G., Allan, J., and Singhal, A. (1996). Automatic text decomposition and structuring. *Information Processing and Management*, 32(2):127 138.

[Sanz et al., 1998] Sanz, I., Berlanga, R., and Aramburu, M. (1998). Gathering metadata from web-based repositories of historical publications. In *9th International Workshop on Database and Expert Systems Applications*, pages 473 478.

[Simpson, 1990] Simpson, P. (1990). *Artificial Neural System*. Pergamon Press.

[Wasserman, 1989] Wasserman, P. (1989). *Neural Computing, Theory and Practice*. Van Nostrand Reinhold.

[Winston, 1992] Winston, P. (1992). *Artificial Intelligence*. Addison Wesley.