OVERVIE

can

h a v e

dimensions

classify or

organize textual pages WWW including

focus

(hyperlink) structure, and various

characteristics the pages including

aim a n d usage

W for avoid e Web paper,

Xerox WWW server. Our goal was to analyze this These a

for extractors will

annotate aggregate WWW pages Web particular,

for classifying

categories, quality

selecting more aware WWW pages

techniques assign

quick prediction

elements WWW pages find display

world, identifying

can techniques rely particular

can extract WWW pages To

Web dimension

WWW pages allowed

specific

pages

the

identif structure strength exhaustive D5 D5 Web walker D5 an

extract

conception Web we t r a v e

use three the Web and The walker pages the

structure. categorization of Web u Hypertext into pages r

a selection pages according their (HTTP) returned g

different of extract pointed Second

sources

text

selection

aggregation Web pages pattern

usage Third

relevancy as measured by activation.

Resource

Locator (Uis R L

and itemsmaintained by the Xerox Web server.

w

requested.

kinds of data were extracted: D5

among

WWW pages at a Web locality.

frequency and *paths* indicate how that

b a and how c e c

many times a traversal was made from one WWW page

WWW pages

locality

These data were used c o o

activation computations respectively:

choose the 1995,

Feature-vector representations of each WWW page

represent values of each page on each dimension.

Graph representations form)

o f from two

breadth-first

was

[illegible] Web sum of the paths to the page was number of

n - s t

w we implemented determine r paths, e performed

timed when the topology

w e The r e disambiguate was consulted

extended. produced

traversals

to determine legitimate while the

s The latter u s p e

relies recently packing s t r a

end-user paths

were created for a machine name when the time between the

determined empirically

users the

boundary paths

created when not

connected to the last page in maintained path. These

maintained New

there recently

r machine for e q

contained d r

number

that identified the frequency of k-substrings for

hereafter creator

computed. generates point

These are WWW pages Web locality

seem for many Entry

points are defined as the set of pages pointed by

are

outside

results

The Xerox

for that machine name, with the ordering the same machine name we children would have average depth of its

two equations of the form:

identified linear

(1)

for nodes X the Web space, where are measured features of each Web and

these were *priori*

us e -1, i may t h

classification for

these for by Linear the machine

learning techniques.

Table the Web pages for order

each of the categories. we hypothesized that

Content Nodes would have few a n d few outlinks,

larger The equation used to

determine this category of Content

size feature, a n d negative inlink

and outlink features. For Head Nodes, we expected the pages

pages have between a n d

children, would more d entry

point based upon actual user navigation patterns.

+1

more sophisticated

[illegible]

O

these the

typical person Web

case, removes typical person page is

identifi

and elsewhere), and **R**

reported purely Iteration spread of

selects

described pages personal project

the

we should sense

Combining Activation Nets

Because

easy spread

weighted p u

different sources and different kinds

and

Consequently,

different directions and

different instance one

identifying pages most

Botafoga have graph-theoretic

These

are based articulation

undirected graph and create of

articulation it

and edges would Botafoga

describe repeat procedure used T

i with n d

number r (nodes

order

functional overconnecting

case, identified

were fact are

important elements of a web group.

graph the

articulation are

size articulation remove

clusters (again g n i r u d least The leads 9

nodes.

web common, often not

extras

web

caused web and a

highly-

book

and sections pointed back

However same nodes. structures are addition, simplified

8

sections web and cluster together number of

was

there these web between