

Web Mining – IN831

Doctorado en Sistemas de Ingeniería

PROFESSORS

Juan D. Velásquez
Víctor Rebolledo L.

Outline

1. Introduction
2. Web Data
3. Knowledge discovery from web data
4. Web Structure Mining
5. Web Content mining
6. Web Usage mining
7. Application

Total = 28 módulos



EVALUATION

- El curso consiste de 2 notas, Tareas (NT), y controles (NC). El cálculo de esas notas se efectúa de la siguiente forma:
- **NC** = Promedio de controles $(C1 + C2 + C3)/3$, donde C_i son notas controles incluido el examen, el cual reemplaza la nota mas baja de los controles si es mayor. El alumno puede eximirse de dar el examen si el promedio actual de controles es de un 5.5 y la nota de tarea (NT) es mayor que 5.5.
- **NT** = Promedio de las entregas de tareas $(T1 + T2 + T3)/3$, donde T_i son las notas de tareas.
- La condición para aprobar el curso es:
- **NC $\geq$ 4.0 y NT $\geq$ 4.0**
- Si no se cumple la condición y las notas se encuentran sobre 3.7, el alumno tiene derecho a un control o tarea recuperativa para optar a nota máxima 4.0.
- La nota final del curso se calcula como:
- **NF = (NT + NC)/2**

Calendario

- ▶ Las actividades serán avisadas en su oportunidad y comprenderán:
 - Controles
 - Laboratorios
 - Tareas

BIBLIOGRAFÍA

- J.D. Velásquez and V. Palade “Adaptative Web Site”. IOS Press, Netherland, 2007.
 - C.D. Manning, P. Raghavan, H. Schutze, “Introduction to Information Retrieval”, Cambridge University Press 2008.
<http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html>
 - G. Myatt, “Making Sense of Data: A practical Guide to exploratory data análisis and data mining”. Wiley Interscience 2007.
 - S. Chakrabarti, “Mining The Web, Discovering Knowlege From HyperText Data”. Morgan Kaufmann Publisher 2003.
 - A. Scime, “Web Mining: Application and techniques.”. IDEA Group Publishing 2005.
- 

Chapter 1

Introduction

PROFESSORS

Juan D. Velásquez
Víctor Rebolledo L.

Summary: Introduction

- The World Wide Web
- E-business
- Toward new portal generation
- Objective of the lecture

The World Wide Web

“The World Wide Web (W3) is the universe of network-accessible information, an embodiment of human knowledge. It is an initiative started at CERN, now with many participants. It has a body of software, and a set of protocols and conventions. W3 uses hypertext and multimedia techniques to make the web easy for anyone to roam, browse, and contribute to”

Tim Berners-Lee (1993)

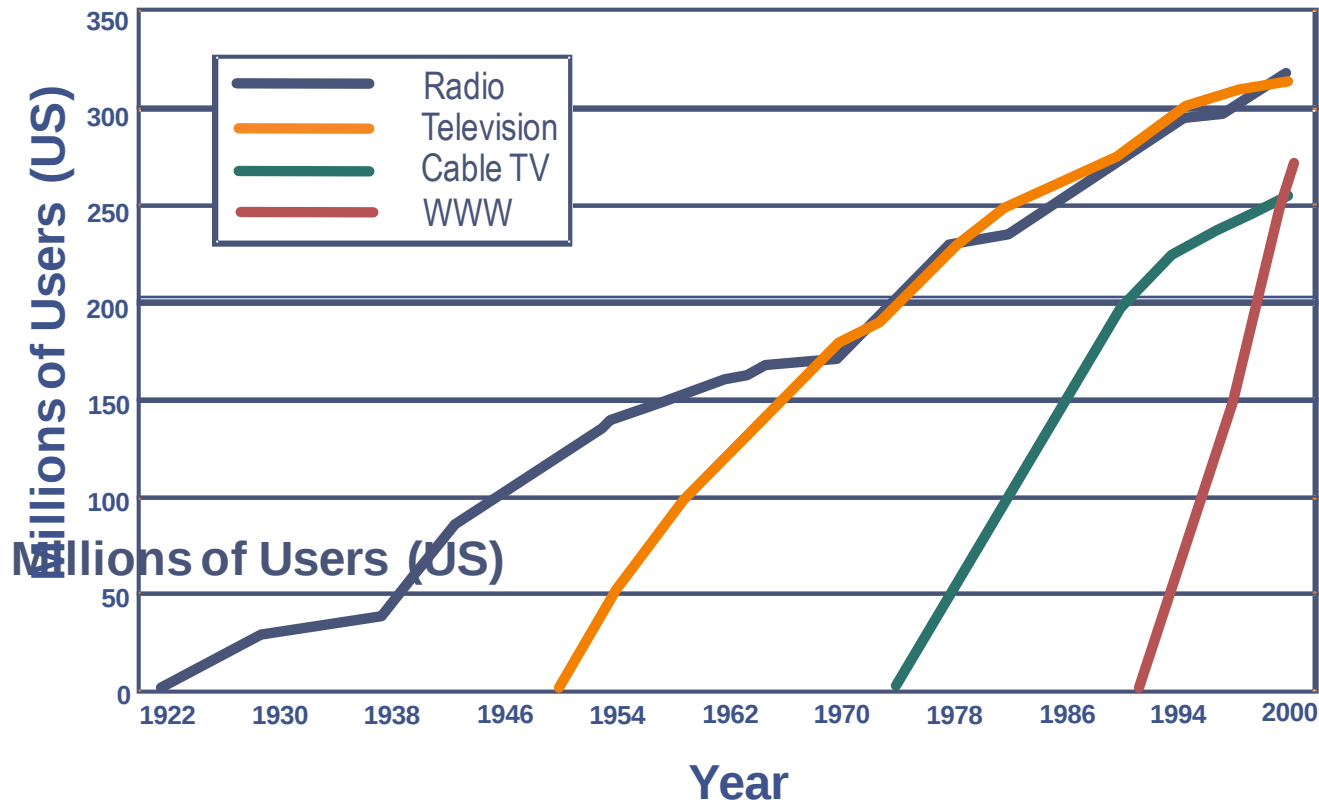


The NET

- ▶ Toward the high speedway Of the information.
- ▶ ¿Who is the owner?
- ▶ **Network of Network**
- ▶ Exponential growth



Learning adoption curve



**Time to reach
200 Million Users:**

Radio 55 Years

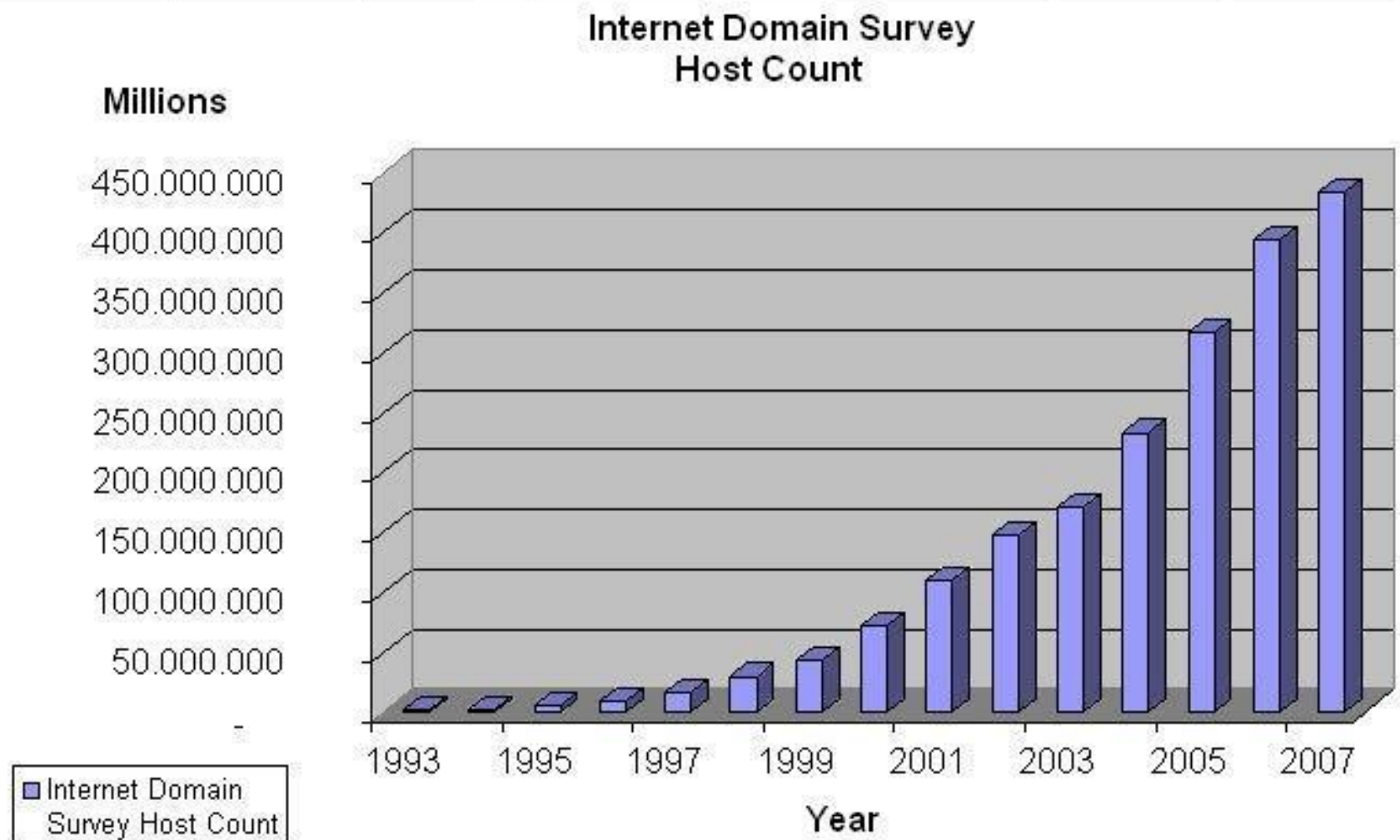
TV 25 Years

Cable ... 12 Years

WWW.... 5 Years

Source: Morgan Stanley Technology Research

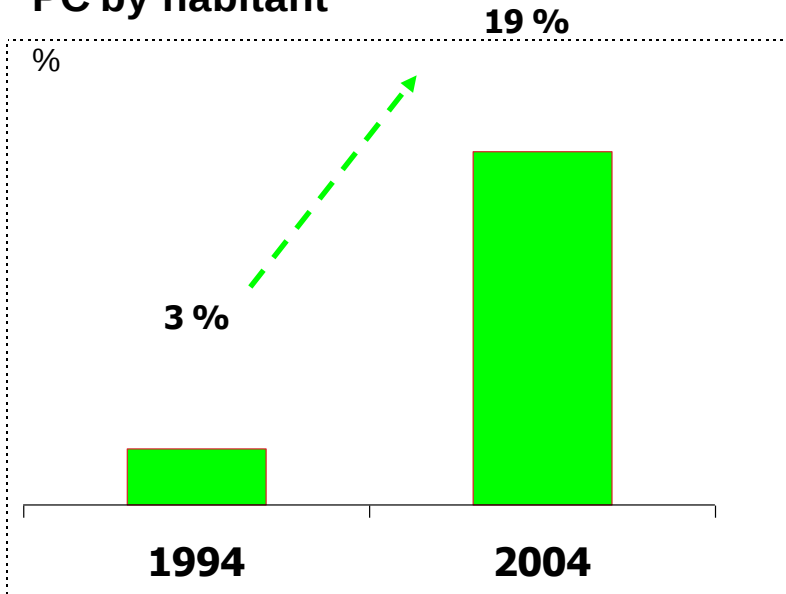
Number of Internet Hosts



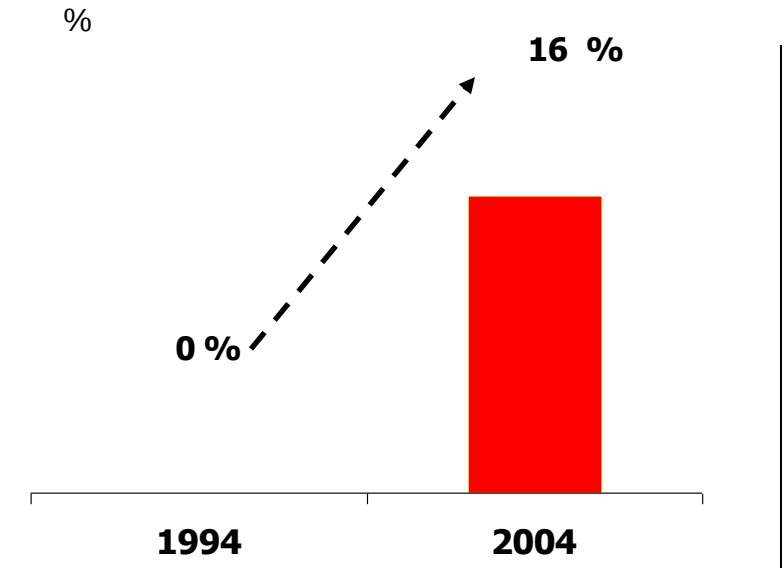
Source: <http://www.isc.org/index.pl?/ops/ds/>

Internet in Chile

PC by habitant



Internet conection Families



Source: Claudio Muñoz, TIN/TIS, 2005

Today 663.000 internet home conections.

Chile

- ▶ 73% of the businesses has an internet conection.
- ▶ 33% of the businesses use Internet as a channel for selling y 59% use electrinic mechanisms electronicos inorder to communicate with otherbusiness.
 - B2B: US\$ 673 millones on 2004
 - B2C: US\$ 80 millones

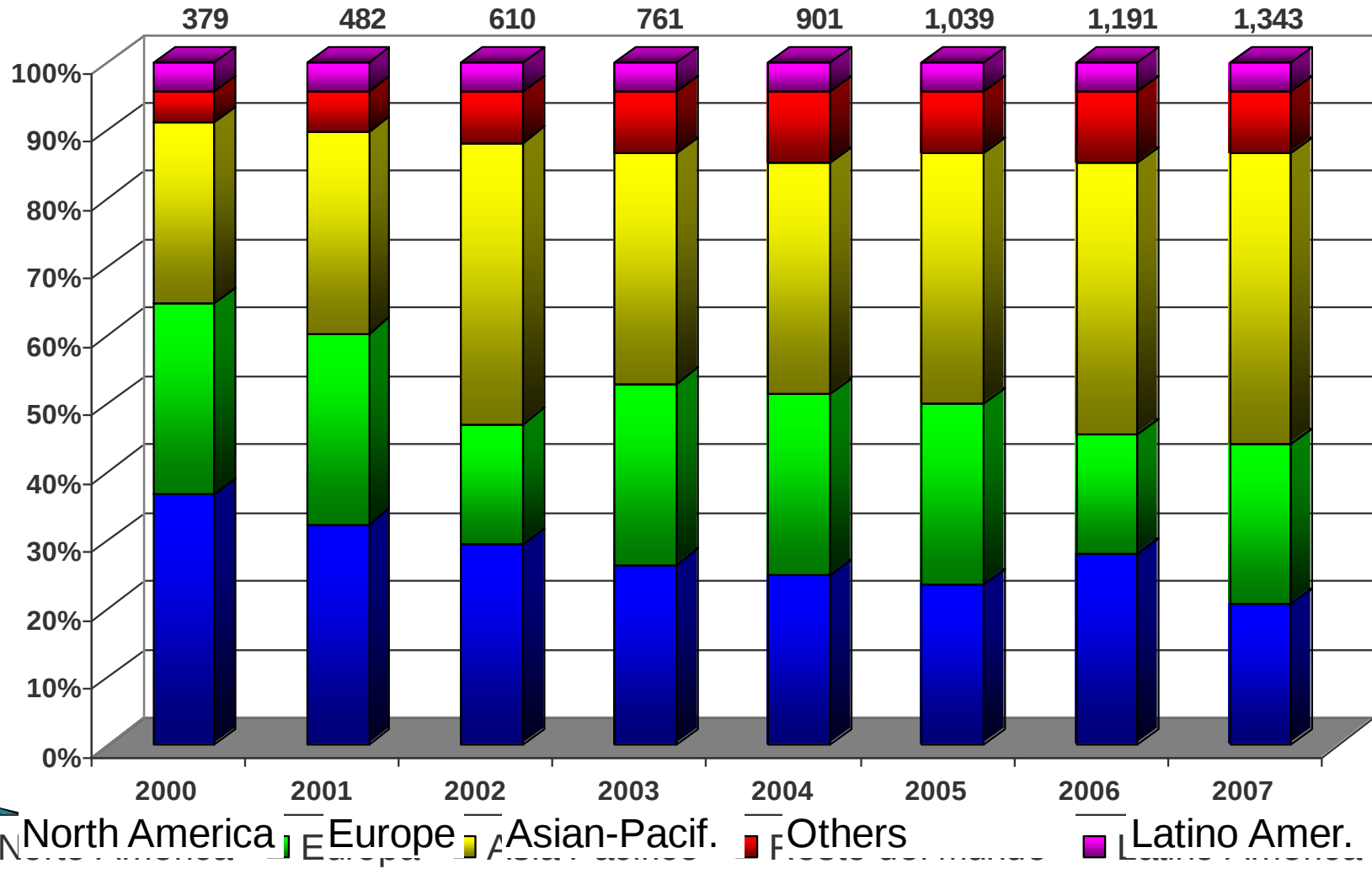
Fuente: El Mercurio 20/08/2006



Internet in the World

Region	Internet User 2004	Internet Users 2007	Annual rate growth (CAGR) 3 Year
Asian Pacific	308MM	588MM	24%
Europe	236	312	10
North America	224	268	6
Rest of the World	87	120	11
Latin America	47	55	6
Total	901MM	1.3B	14%

Geographic Distribution of Internet Users



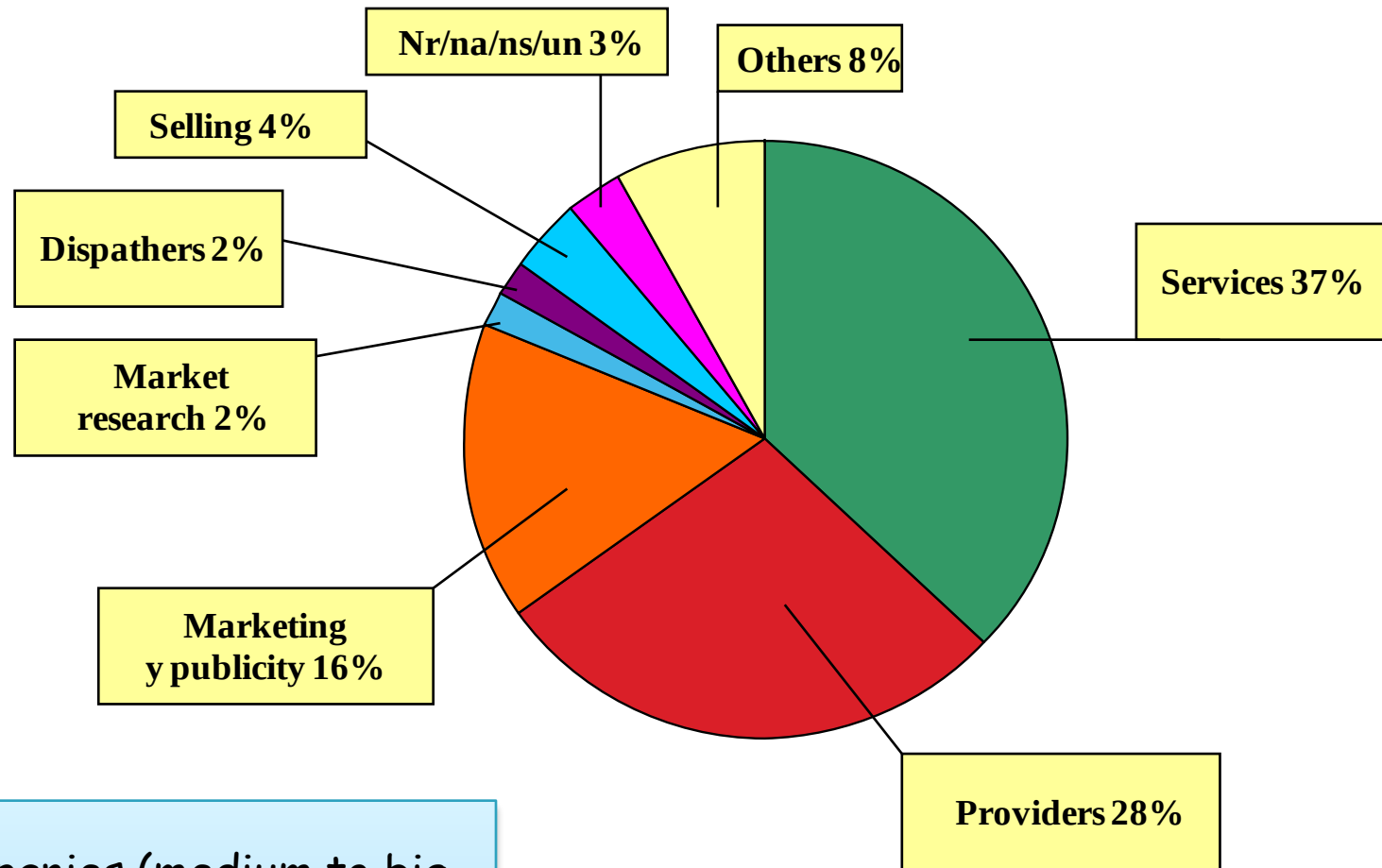
The computer is the network, the network is the computer

- The web is changing everything.
 - The new economy: Google Model, Amazon, e-Banking.
 - Change in the Supply Chain.
 - Consumer directly perform OnLine product request.
 - Reducing Information asymmetry gap.
 - New kind of problem: Retain consumer in a web environment, The new web consumer profile, the web site as the e-commerce “service”.
- 

E – Business

- ▶ Business to Business (B2B)
 - ▶ Business to Consumer (B2C)
 - ▶ Peer to Peer (P2P)
- 

E-business distribution



- 134 companies (medium to big ones).
- 48% with 100 to 500 employees
- 52% over 500 employees

Source: IDC

E-business

- ▶ Something change abruptly.
 - ▶ Supply chain was altered.
 - ▶ If an Intermediary doesn't add value to products, then they will be erased.
 - ▶ The new business model based on the new internet channel, lowers the chain costs.
- 

E-business (2)

► Transactions Cost

Business	Transaction type	Cost (US\$)	
Costo by transaction			
	Cashier	1,07	
	Phone	0,52	
	ACM	0,27	
	Internet	0,13	
Cost of plane ticket			
	Travel Agency	8,0	
	Internet	1,0	
Insurance			
	Agent	550	
	Internet	275	
Software			
	Reseller	15	
	internet	0,35	

But ... there are some warnings



NASDAQ index evolution dotcom

The web portal: Our Point of Sale

- What is the **ideal structure and content** of a web site?
- Different users have distinct goals
 - The behaviour of users changes over time.
 - Sites must be restructured as they grow to meet current needs, typically by accumulating pages and links.
- 

The Adaptive Web Site (AWS)

- ▶ Based on **user behaviour**
- ▶ Web Site **recommendation**
- ▶ Use of Web Intelligence (WI)
 - <http://wi.dii.uchile.cl>
- ▶ Understanding **user preferences**
- ▶ **Applications**
 - ▶ Web Usage Mining
 - ▶ Web Structure Mining
 - ▶ Web Content Mining
- ▶ Use of **Information Retrieval**

Objectives of this lecture

- ▶ To know about **data mining techniques on the web.**
 - ▶ Why **web mining** is important in the **e-business world.**
 - ▶ How to perform a **web mining project**
 - ▶ What are the **different application fields** of web mining
- 

Data Mining techniques on the Web and benefits

- ▶ *Web Intelligence (WI)*
 - ▶ Web Data: Very large amount of
 - Logs
 - Text and multimedia content
 - Structure of links
 - ▶ Several tools of data mining apply to this field:
 - Clustering, regressions, association rules.
 - ▶ Important benefits returns from the mining process:
 - Google growth, e-business, e-market campaigns, CRM applications.
- 

How to perform webmining projects

- ▶ **Learning by doing**
- ▶ **3 programming projects** to be performed by student during the semester.
 - **Web Structure Mining:**
 - Mini Crawler.
 - **Web Usage Mining.**
 - **Web Content Mining.**

Applications of web mining

- ▶ Recommendation System
 - ▶ System for personalization
 - ▶ Web Personalization
 - ▶ Adaptive Web-based system
- 