Produced by:
Dr. Brenda Mullally bmullally@wit.ie
Ruth Barry rbarry@wit.ie

Department Computing Maths and Physics
Waterford Institute of Technology
www.wit.ie
moodle.wit.ie

MSc Enterprise Software Systems
Business Intelligence

Information Integration

Information Integration

- How is information integrated?
  - Synthesis of new insights from unstructured data residing in the organisation's enterprise systems, such as enterprise portals and document management systems.
  - Creation of new insights via the integration of structured organizational data with external data, such as Web-based unstructured information from customer Web sites or vendor data sources.

Why integrate with unstructured data?

Environmental Scanning

- Scanning for information about events and relationships in a company's outside environment,
  - the knowledge of which would assist top management in its task of planning the company's future course of action
- Improve organizational performance
- 'looking for' and 'looking at' information

Environmental Scanning

- Brown and Weiner (1985) define environmental scanning as "a kind of radar to scan the world systematically and signal the new, the unexpected, the major and the minor"

Environmental Scanning-Uncertainty

- Industry or market
- Technology
- Regulatory
- Economic
- Social
- Political

How do we scan?

- Techniques for identifying and extracting external business information.
  - Text mining
  - Web mining

Text Mining

- According to a study by Merrill Lynch & Gartner, 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text).
- Unstructured corporate data is doubling in size every 18 months.
- Tapping into these information need to stay competitive.

Text Mining

- A semi-automated process of extracting knowledge from unstructured data sources
- Data source may not reside in a structured database but is more likely in an unstructured file.

Text Mining

- Algorithms focus on:
  - Information retrieval
  - Information extraction
  - Information summarization

- Document text mining relies on text categorization (TC) techniques to uncover new knowledge from it.
- IR indexing

Natural Language Processing (NLP)

- Structuring a collection of text
  - Old approach: bag-of-words
  - New approach: natural language processing
- NLP is
  - a very important concept in text mining.
  - a subfield of artificial intelligence and computational linguistics.
  - the study of "understanding" the natural human language.
- Syntax versus semantics based text mining

Natural Language Processing (NLP)

- What is "Understanding" ?
  - Human understands, what about computers?
  - Natural language is vague, context driven
  - True understanding requires extensive knowledge of a topic

  - Can/will computers ever understand natural language the same/accurate way we do?

Natural Language Processing (NLP)

- Challenges in NLP
  - Part-of-speech tagging
  - Text segmentation
  - Word sense disambiguation
  - Syntax ambiguity
  - Imperfect or irregular input
  - Speech acts
- Dream of AI community
  - to have algorithms that are capable of automatically reading and obtaining knowledge from text

Examples

- The professor said on Monday he would give an exam.
- The chicken is ready to eat.
- Visiting relatives can be boring.
- "A lady with a clipboard stopped me in the street the other day. She said, 'Can you spare a few minutes for cancer research?' I said, 'All right, but we're not going to get much done.'"
  (English comedian Jimmy Carr)
- They are cooking apples.
- “cold” disease, temperature sensation, environmental condition?

Natural Language Processing (NLP)

- WordNet
  - A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets
  - A major resource for NLP
- Sentiment Analysis
  - A technique used to detect favorable and unfavorable opinions toward specific products and services, e.g. customer feedback

Text Mining Concepts

- Benefits of text mining are obvious, especially in text-rich data environments
  - e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.
- Electronic communication records (e.g., Email)
  - Spam filtering
  - Email prioritization and categorization
  - Automatic response generation

Text Mining Application Area

- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Concept linking
- Question answering

Text Mining Applications

- Marketing applications
  - Enables better CRM
- Security applications
  - ECHELON, OASIS
  - Deception detection (…)
- Medicine and biology
  - Literature-based gene identification (…)
- Academic applications
  - Research stream analysis

Text Mining Tools

- Commercial Software Tools
  - SPSS PASW Text Miner
  - SAS Enterprise Miner
  - Statistica Data Miner
  - ClearForest
- Free Software Tools
  - RapidMiner
  - GATE
  - Spy-EM

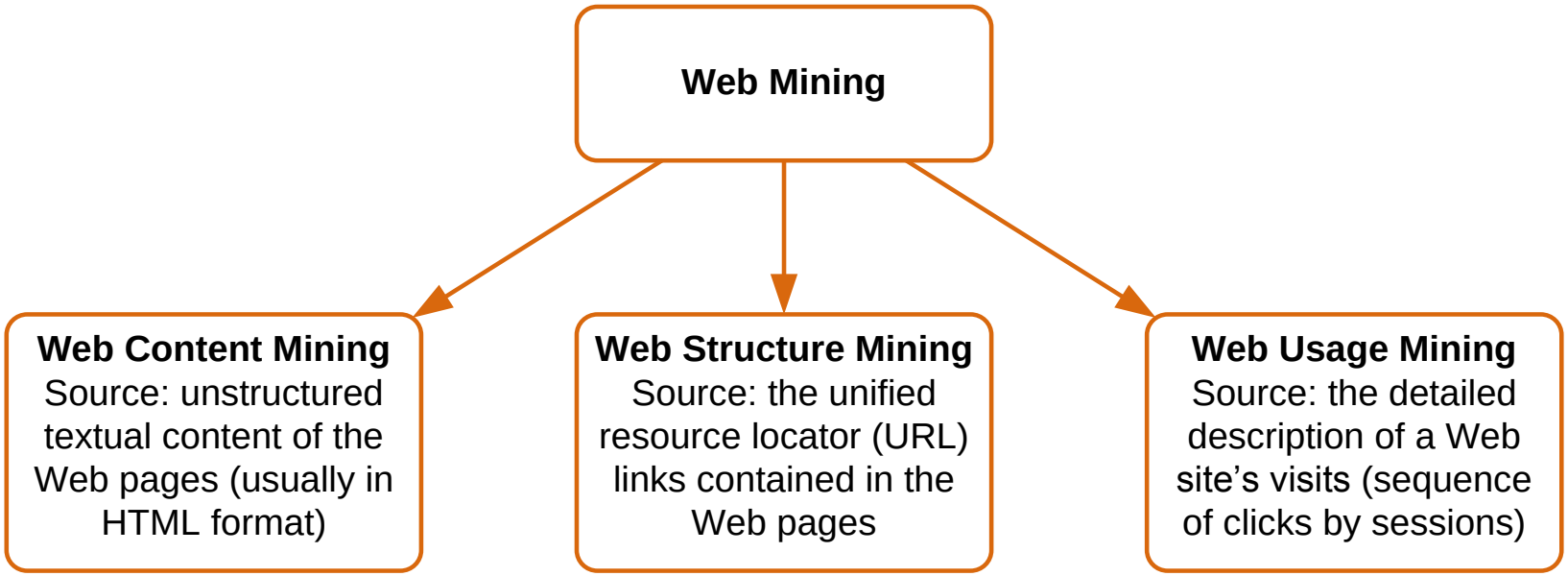
Web Mining Overview

- Web is the largest repository of data
- Data is in HTML, XML, text format
- Challenges (of processing Web data)
  - The Web is too big for effective data mining
  - The Web is too complex
  - The Web is too dynamic
  - The Web is not specific to a domain
  - The Web has everything
- Opportunities and challenges are great!

Web Mining

- Web crawling with on-line text mining
- Several differences between traditional data mining and web mining are:
  - Web mining requires linguistic analysis abilities.
  - Web mining requires techniques from both information retrieval and artificial intelligence domains.
  - Web pages are indexed by the words they contain, using information retrieval (IR) techniques.
  - Web content mining relies on text categorization (TC) techniques.

Web Mining - Classified



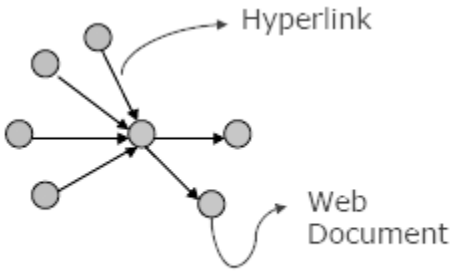
Web Mining- Uses

- There are three types of uses for web mining:
  - *Web structure mining*
  - *Web usage mining*
  - *Web content mining*

Web Structure Mining

- Examination of how documents on the Web are structured, seeking to discover the model underlying the Web link structures.
- Intra-page structure mining
- Inter-page structure mining

- http://webdatacommons.org/hyperlinkgraph
- http://law.di.unimi.it/index.php



Web Graph Structure

Web Content Mining

- Textual content on the web

- Web content data includes analysis of the semi-structured and unstructured content used to create the Web page, which includes the text, images, audio, video, hyperlinks, and metadata.
- Web content mining is based on text mining and IR techniques.

- http://commoncrawl.org

Google – web mining application

- Caffeine algorithm released in 2010 designed to improve indexing speed and provide users with fresher results.
- Hummingbird released in 2013 deigned to understand concepts, the relationships between concepts and more complex questions.

- http://searchengineland.com/googles-new-indexing-infrastructure-caffeine-now-live-43891

- http://moz.com/google-algorithm-change

Web Usage Mining

- Also called Web Analytics
- Extraction of information from data generated through Web page visits and transactions
  - data stored in server access logs, referrer logs, agent logs, and client-side cookies
  - user characteristics and usage profiles
  - metadata, such as page attributes, content attributes, and usage data
- Clickstream data
- Clickstream analysis

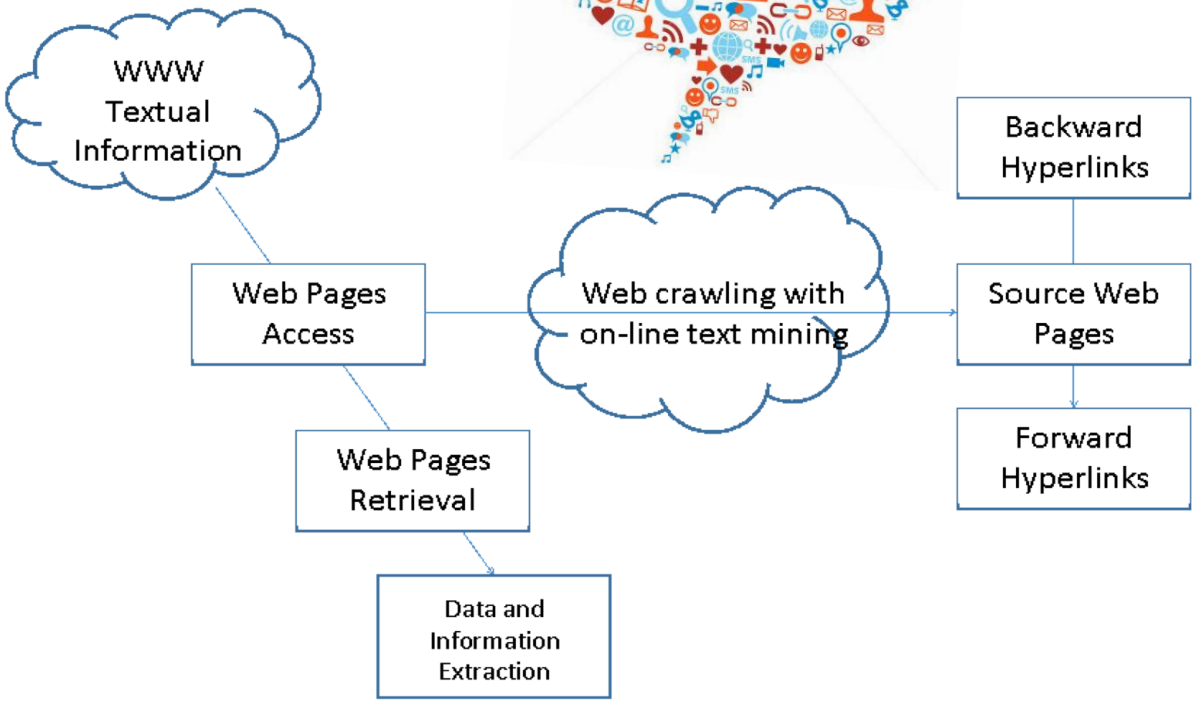
Web Usage/Web Analytics mining

- Web analytics programs can document a marketing campaign or manage the efforts of products and services
- There are four categories of metrics that can directly impact a business objectives:
  - Web site usability: How were they using my web site?
  - Traffic sources: Where did they come from?
  - Visitor profiles: What do my visitors look like?
  - Conversion statistics: What does it all mean for the business?

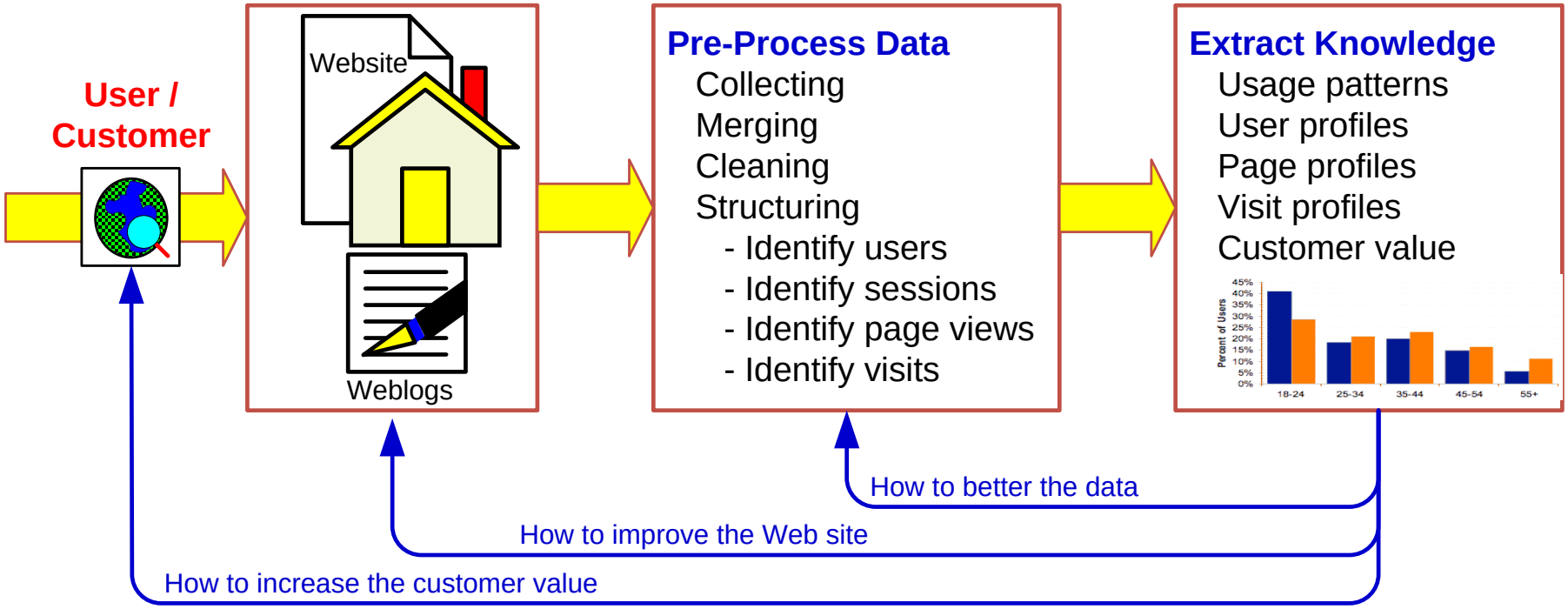
Web Usage Mining

- Web usage mining applications
  - Determine the lifetime value of clients
  - Design cross-marketing strategies across products.
  - Evaluate promotional campaigns
  - Target electronic ads and coupons at user groups based on user access patterns
  - Predict user behavior based on previously learned rules and users' profiles
  - Present dynamic information to users based on their interests and profiles

Web Mining Process



Web Usage Mining (clickstream analysis)



Mining Web Data-Advantages

- Developing a personalised relationship with online customers
- Improving the profitability of online stores through improved processes
- Growing online revenues

Application of Web Mining in Customer Relationship Management (CRM)

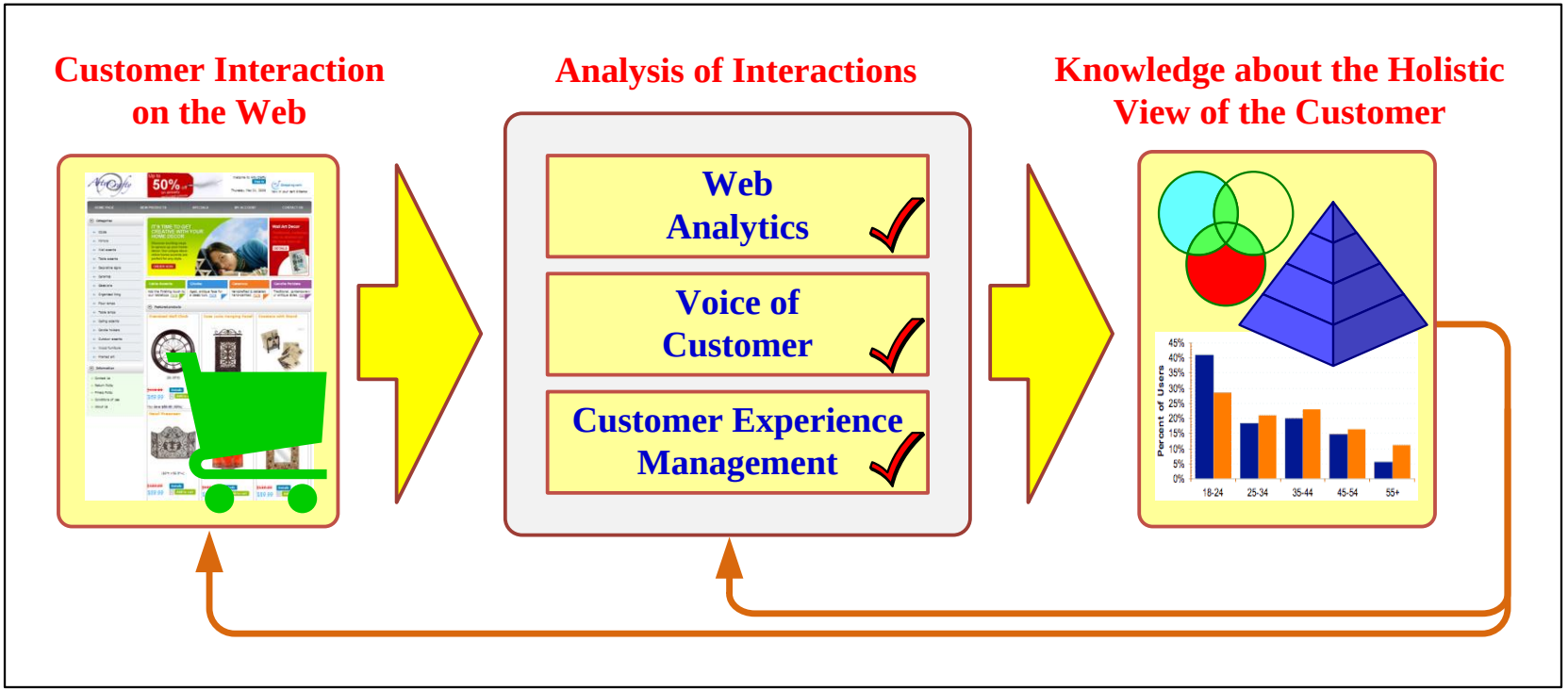
- Applications in CRM can be characterized as being:
  - Operational CRM
  - Analytical CRM

CRM Adoption Purpose

- Integrate the customer viewpoint across all aspects
- Respond to customer demands in "Internet time"
- Gain more value from BI investments

Web Mining Success Stories

- Amazon, Google, Facebook
- Website Optimization Ecosystem



WebLinks

- http://webtrends.com/
- http://miningthesocialweb.com/
- http://www.technologyreview.com/view/527746/how-advanced-socialbots-have-infiltrated-twitter/

Web Mining Tools

Product Name	URL
Angoss Knowledge WebMiner	angoss.com
ClickTracks	clicktracks.com
LiveStats from DeepMetrix	deepmetrix.com
Megaputer WebAnalyst	megaputer.com
MicroStrategy Web Traffic Analysis	microstrategy.com
SAS Web Analytics	sas.com
SPSS Web Mining for Clementine	spss.com
WebTrends	webtrends.com
XML Miner	scientio.com