Produced by:

Ruth Barry

Dr. Brenda Mullally

rbarry@wit.ie

bmullally@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.

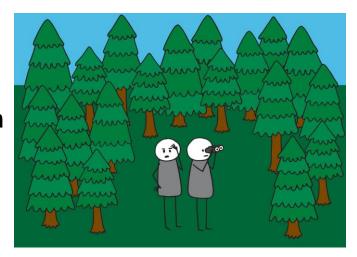




Why integrate with unstructured data?

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"
- "Searching the environment for important events or issues that might affect an organisation"





Environmental Scanning-Uncertainty

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





How do we scan?

▶ Technologies and techniques for identifying and extracting external business information.

- Text mining
- Web mining





Text Mining

- According to a study by Merrill Lynch & Gartner, 85% 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 sources is not an option, but a need to stay competitive.
- Answer: text mining
 - A semi-automated process of extracting knowledge from unstructured data sources
 - a.k.a. text data mining or knowledge discovery in textual databases



Sharda, et. al. (2017) Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Ed., Ch 5.



Text Analytics and Text Mining

- Text Analytics versus Text Mining
- Text Analytics =
 - Information Retrieval +
 - Information Extraction +
 - Data Mining +
 - Web Mining

or simply

Text Analytics = Information Retrieval + Text Mining

Slide 5-8

Text Mining

- Applying software technology to understand volumes of (unstructured) text.
- Analyzing the data to determine which terms are more prevalent than others.
- Learning how terms and phrases are related to one another.

Understanding what the common themes in the document collection are.
automated data mining survey

er transcripts

oot cause

responses co qualatativ

customer dashboards consume

trends ad-hoc analysis early warning

classificati**(t)** ad-hoc an reviews se

https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/discovering-what-you-want-107347.pdf



Text Mining Concepts

- Benefits of text mining are obvious especially in textrich 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 communization records (e.g., e-mail)
 - Spam filtering
 - ▶ E-mail prioritization and categorization
 - Automatic response generation

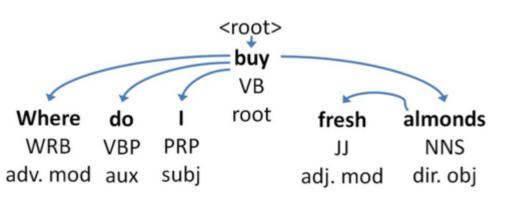


Text Mining Application Area

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

Text Mining Terminology

- Unstructured data
- Corpus (and corpora)
- Terms
- Concepts
- Stemming
- Stop words (and include words)
- Synonyms (and polysemes)
- Tokenizing



Text Mining Terminology

- Term dictionary
- Word frequency
- Part-of-speech tagging
- Morphology
- gging Where do I root fresh almonds
 WRB VBP PRP
 adv. mod aux subj adj. mod dir. obj

<root>

- Term-by-document matrix
 - Occurrence matrix
- Singular value decomposition
 - Latent semantic indexing

POS Tags and their meanings

Tag	Description
СС	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

_	
Tag	Description
PRP\$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
ТО	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP\$	Possessive whpronoun
WRB	Whadverb

http://blog.thedigitalgroup.com/sagarg/2015/06/18/part-of-speech-tagging-using-opennlp/



Text Mining algorithms

1

Focus on Meaning

Identify parts of speech, identify sentiment, and use meaning of words to analyze text 2

Bag of Words

Use methods that treat words simply as tokens of distinct categories without understanding meaning



Natural Language Processing (NLP)

- Structuring a collection of text
 - Old approach: bag-of-words
 - New approach: natural language processing

What is NLP?

Wiki definition: Natural language processing (NLP) is a field o computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.

Goal

- For computers to process or "understand" natural language in order to preform tasks that are useful e.g.
 - Performing tasks, like making appointments, buying things
 - □ Question answering Siri, Google assistant, Facebook M,
- Full understanding and representation of language





Natural Language Processing (NLP)

Challenges in NLP

- Part-of-speech tagging
- Text segmentation
- Word sense disambiguation
- Syntax ambiguity
- Imperfect or irregular input
- Speech acts

context personal vowels know stempt tends to send the personal works working to send the personal works working the personal works working

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
- Need automation to be completed

Sentiment Analysis

- A technique used to detect favorable and unfavorable opinions toward specific products and services
- SentiWordNet



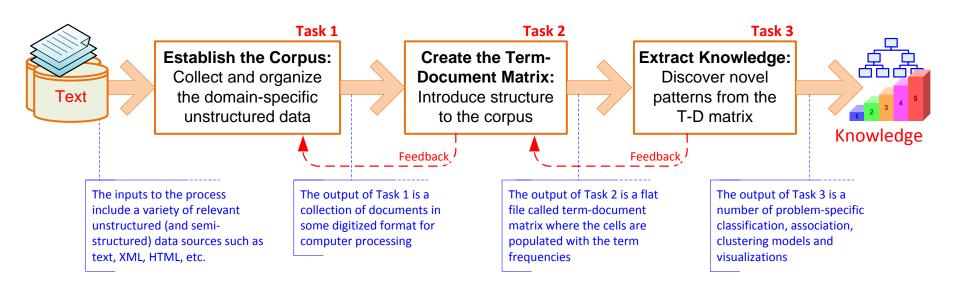
NLP Task Categories

- Question answering
- Automatic summarization
- Natural language generation & understanding
- Machine translation
- Foreign language reading & writing
- Speech recognition
- Text proofing, optical character recognition
- Optical character recognition



Text Mining Process

The Three-Step/Task Text Mining Process



Sentiment Analysis

- ▶ Sentiment → belief, view, opinion, and conviction
- Sentiment analysis is trying to answer the question "What do people feel about a certain topic?"
- By analyzing data related to opinions of many using a variety of automated tools
- Used in variety of domains, but its applications in CRM are especially noteworthy (which related to customers/consumers' opinions)





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!

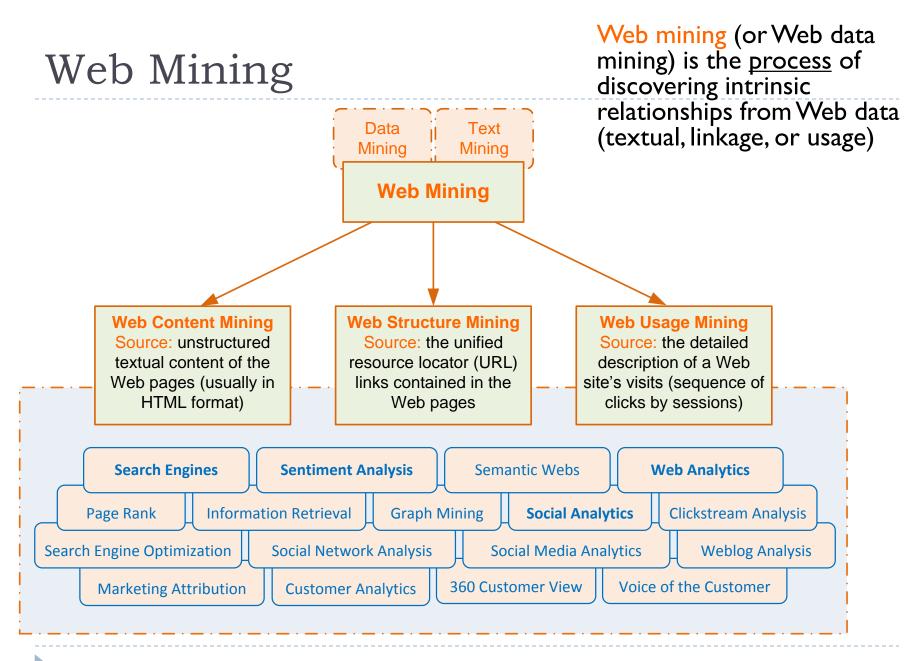




Web Mining- Uses

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





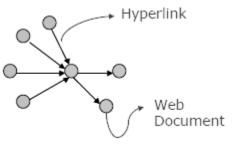
Web Content/Structure Mining

- Mining the textual content on the Web
- Data collection via Web crawlers



- Web pages include hyperlinks
 - Authoritative pages
 - Hubs
 - Hyperlink-induced topic search (HITS) alg.

Slide 5-28



Web Graph Structure

Search Engines

- A "search engine" is a key example for information retrieval systems, concepts and techniques in text mining.
- Search engine is a software program that searches for documents (Internet sites or files) based on the keywords (individual words, multi-word terms, or a complete sentence) that users have provided that have to do with the subject of their inquiry

Anatomy of a Search Engine

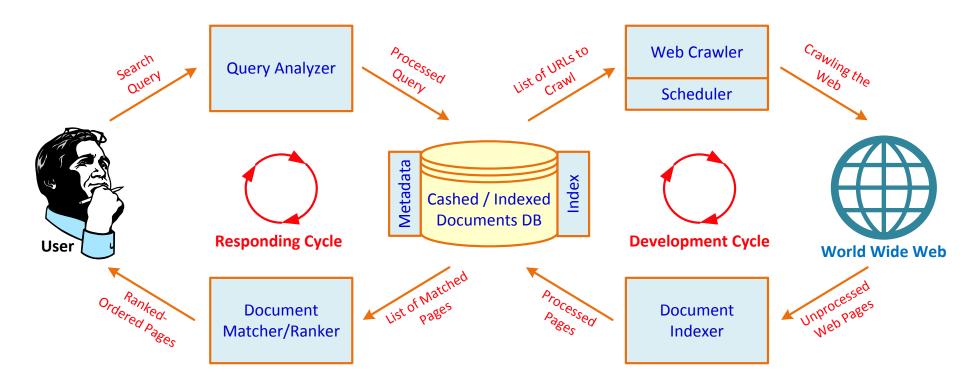
Development Cycle

- Web Crawler
- Document Indexer

2. Response Cycle

- Query Analyzer
- Document Matcher/Ranker

Structure of a Typical Internet Search Engine



Web Usage Mining

- 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

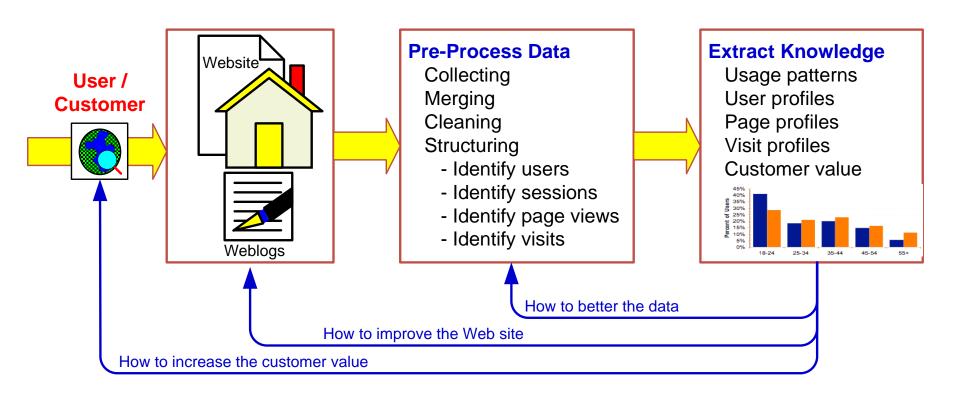


Web Usage Mining

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 Usage Mining (Clickstream Analysis)



Web Analytics Metrics

- Web analytics programs can document a marketing campaign or manage the efforts of products and services
- Web site usability
 - How were the visitors 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 Analytics Metrics

Web Site Usability

- Page views
- Time on site
- Downloads
- Click map
- Click paths

Traffic Source

- Referral Web sites
- Search engines
- Direct
- Offline campaigns
- Online campaigns



Web Analytics Metrics

Visitor Profiles

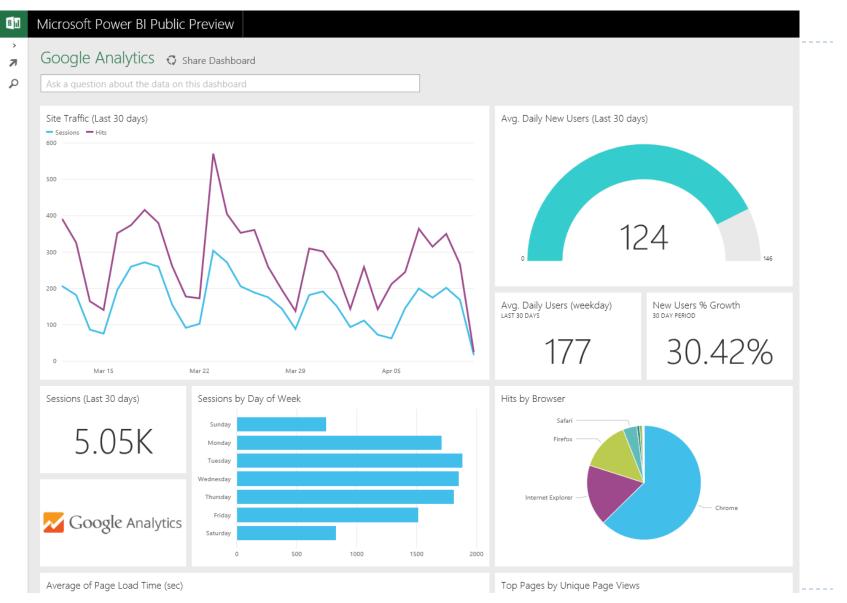
- Keywords
- Content groupings
- Geography
- Time of day
- Landing page profiles

Conversion Statistics

- New visitors
- Returning visitors
- Leads
- Sales/conversions
- Abandonment/exit rate



A Sample Web Analytics Dashboard



Web Mining Success Stories

- Amazon, Google, Facebook
- Website Optimization Ecosystem

