

***An analysis of web sites as a
communication tool:
an application in the banking sector***

Silvia Biffignandi

Bibliography

Datamining come approccio alle analisi dei mercati e delle performance aziendali, Silvia Biffignandi editor, Quaderni del Dipartimento di Matematica, Statistica, Informatica e Applicazioni, n.8/03, 2003

author and title of the paper:

S.Biffignandi et al., Web communication behaviour of banks: a case study

University reseach grant

To be published F. Angeli, Milano

S. Biffignandi, An analysis of web sites as a communication tool

Outline of the talk

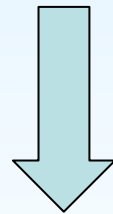
- ✓ Aim of the study
- ✓ The technique
- ✓ The data set and step of analysis
- ✓ Results

Aim of the study

- ✓ highlight the communication strategies via Internet (the Web) of some banking institutes which belong to the main financial groups operating in Italy.

Why ?

- ✓ Financial market: significant changes during the 1990s



Changes in strategies

Why websites?

the use home websites to communicate with their customer base

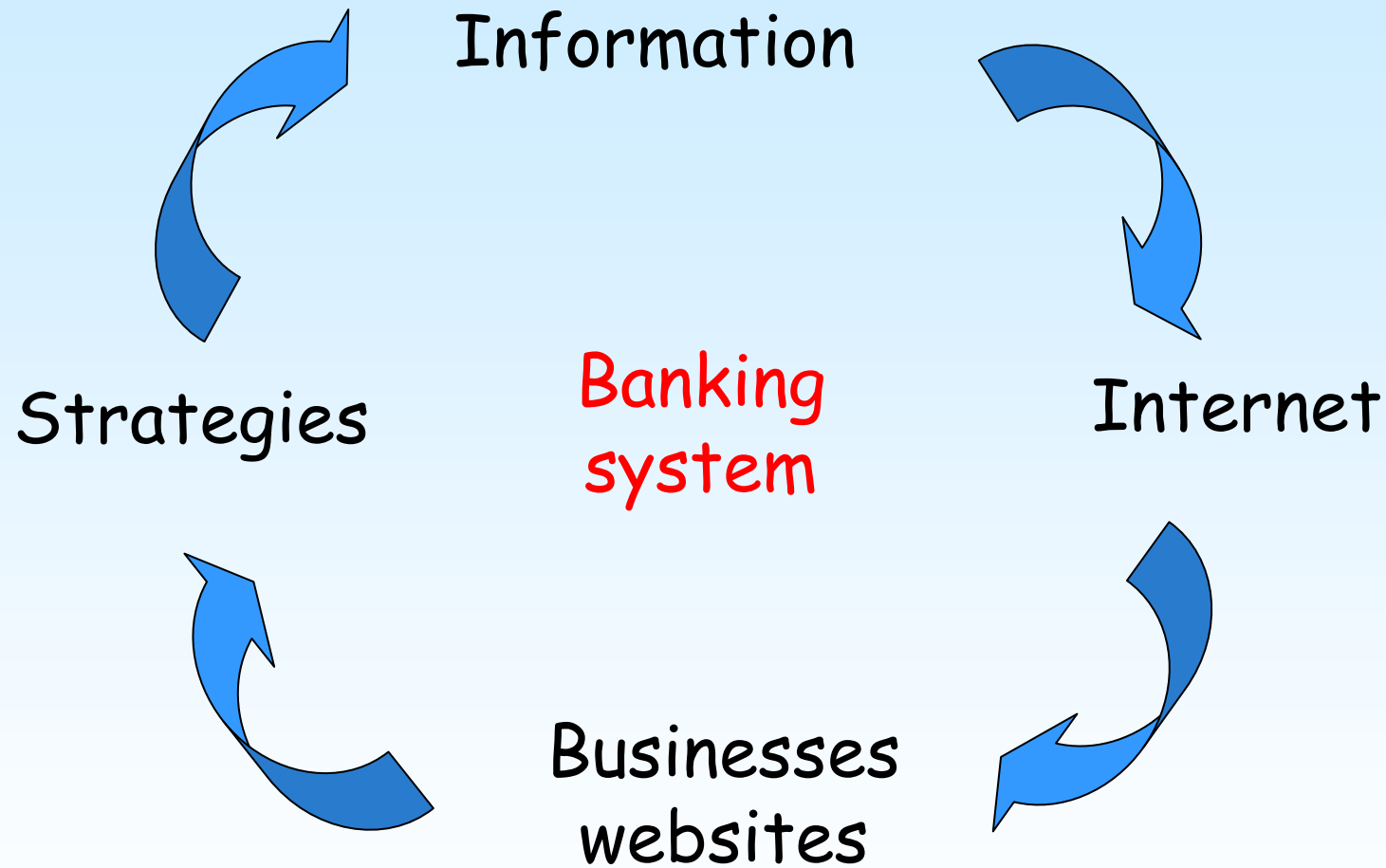
defines an extremely interesting element used to examine

both **how coherent** the communication lay out is with the services offered by the bank and with its target market as well as being used as **a reference model** for the construction of other web sites.

Specific tasks?

a statistical study on the content of banking institution web sites

mainly to determine the characteristics of each site, to evaluate how the institutes exploit the web and to identify different behavioural patterns.



S. Biffignandi, An analysis of web sites as a communication tool

The technique?

TEXT MINING



sites as a communication tool

Data mining, text mining, text and content analysis,
web mining??

Some concepts

KDD/Data Mining

Knowledge Discovery in Databases (or KDD) through data mining is the process of “extracting previously unknown, valid, and actionable knowledge from large databases” and then of using it to make crucial business decisions.

(W. Frawley et al., Knowledge Discovery in Databases: An Overview. AI Magazine , 1992)

KDD/Data Mining

"The science of extracting useful information from large data sets or databases" .

D. Hand et al. Principles of Data Mining, 2001

KDD/Data Mining

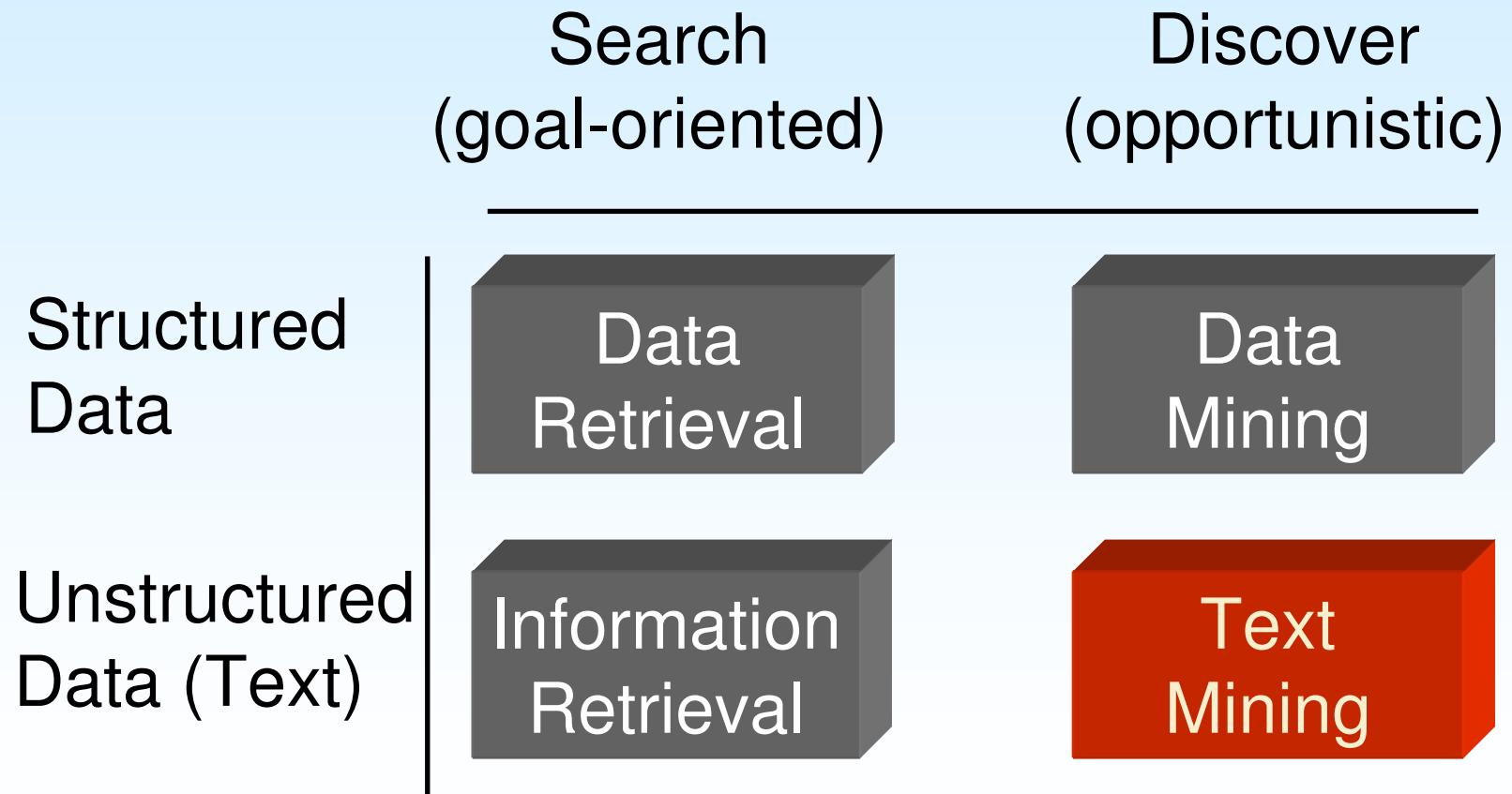
The mined data are generally **structured data**, regarding, very often, marketing problems.

Text Mining

*a particular form of data mining
where
the **data are in textual format.***

*Seen that the data are in an **unstructured format,**
the data preparation step is longer than usual
and requires a linguistic preprocessing
to solve (also if partially) the ambiguities.*

“Search” versus “Discover”



S. Biffignandi, An analysis of web sites as a communication tool

*As in **data mining**
the data consist
in subjects (e.g. customers)
defined by a **set of features** (e.g. age, salary,
accounts...),
in **text mining** t
he data consists in
documents defined by a **set of words.***

Principal difference

in data mining,
**the features that characterize a subject are
tens, or hundreds, with a limited
variability,**

in text mining,
**the words that characterize a document
are thousands, with a variability of
hundreds of thousands.**

Principal difference

The feature selection in the case of text mining needs particular attention and elaboration, with help coming from automatic filtering and/or human intervention.

Text and content Analysis

text analysis is more than text mining,

- ***it may or may not include "text mining",***
- ***it is more to do with trying to understand how human brain analyses text and trying to automate those processes in a manner that may not involve some very complicated AI based technology.***

Web mining

Internet is becoming one of the main media by which documents, data and information can be obtained for client information targets

- ❑ the data sets in Web Mining are extremely large
- ❑ main advantage is that this flood of information is often free
- ❑ main disadvantage is that generally this information is too much, and it is not easy to detect where the important piece of information is located.

Web analysis

***The key component of Web mining
is the mining process itself***

.....

***as well as developing some
techniques of its own,***

e.g. path analysis.

Web analysis

It can be said to have three operations of interests:

- ❑ clustering (finding natural groupings of users, pages etc.),***
- ❑ associations (which URLs tend to be requested together)***
- ❑ sequential analysis (the order in which URLs tend to be accessed).***

Web analysis

Issues on server site data collection

- ❑ integrating various data sources (*server access logs, referrer logs, user registration or profile information*);
- ❑ difficulties in the identification of users due to missing unique key attributes in collected data;
- ❑ identifying user sessions or transactions from usage data, site topologies, and models of user behavior.

S. Biffignandi, An analysis of web sites as a communication tool

Web analysis

has been used in two distinct ways:

1)the first, which is referred to as ***Web content mining***, describes the process of information or resource discovery from millions of sources across the World Wide Web.

Web analysis

2) The second (***Web usage mining***) is the process of mining Web access logs or other user information user browsing and access patterns on one or more Web localities.

The case study

S. Biffignandi, An analysis of web sites as a communication tool

The data set

- ✓ the “About Us” (“Chi siamo”) presentation pages of 28 banking institutes
- ✓ in addition, the relative links on two levels.

BANCA ANTONVENETA Group	ANTONVENETA
BNL Group	BNL
BANCA CARIGE Group	CARIGE
BANCA LOMBARDA Group	Banca Lombarda, Banco di Brescia , Brebanca
CREDITO EMILIANO Group	Credem

UNICREDITO Group	Xelion, Unicredito Unicreditbanca , Unicreditprivate Unicreditimpresa
CAPITALIA Group	Irfis, MCC, Banca della Rete, Bipop, Capitalia
SAN PAOLO IMI Group	San Paolo, Banco di Napoli, Fideuram, Farbanca, Finemiro
GLOBAL VALUE Group	Banca Intesa , Bancacis, Carifol, Carispo
BANCA CR FIRENZE Group	Banca crfirenze
CARISMI Group	Banca Carismi S. Biffighandi, An analysis of web sites as a communication tool

The technique?

TEXT MINING



sites as a communication tool

Step of analysis

- ✓ linguistic analysis
- ✓ data reduction
- ✓ clustering

Step of analysis: Linguistic

Unstructured documents

the unstructured textual material is transformed into a format that can be analyzed by modeling.

The objective of this step is to extract automatically the features from the document, that is to recognize and classify the items present in natural language texts.

Step of analysis: Linguistic

This objective is reached through **linguistic analysis**, whose the three basic components are:

1. a **language model**, offered in principal languages, that provides the grammatical knowledge to break sentences into their basic components including nouns, verbs, adjectives, dates etc. selecting the new keywords from the title or the abstract or the total document.
2. a **generic dictionary** of words and multi-word expression, which can be enhanced with industry-specific dictionaries to support the process of categorizing highly technical documents.

Step of analysis: Linguistic

3. a “**Relations Extraction**” engine. All the sentences where there are relations as “subject X” “verb or expression meaning” “subject Y” are detected. This component allows the detection of the so-called “signals” [e.g., alliances].

Step of analysis: Linguistic

- ✓ **Linguistic step** (by using dictionaries, syntactic taggers, lemmatization engines, search engines):
 - *solves the ambiguities linked to the language* (for instance, the words:
 - - ‘record’ in the English sentence “we record this record”)
 - *recognizes the semantic value of the words* (electronic dictionaries are, with synonym maps, for the languages we want to manage),
 - *lemmatizes the words* (“International Business Machines” is transformed in “IBM”; “loves”, “love”, “loved” are transformed in “to love”),
 - *automatically indexes* (i.e. associates the key concepts/best indexes to the chosen document).

Step of analysis: Linguistic

- Thesaurus construction
- Using co-occurrence statistics to detect semantic regularities across concepts
- Related to Latent Semantic Indexing and Concept

Step of analysis: Linguistic

- Latent Semantic Indexing

- It has been presented as a way to capture the main semantic dimensions in a text collection, avoiding synonymy and polysemy problems
- exploits co-occurrence information between concepts to derive a text representation based on new, less dimensions
- can be seen as an effective dimensionality reduction method

Berry et al., 1995

Step of analysis: data reduction

Latent Semantic Indexing

- The basic idea is mapping a high-dimensional space into a low-dimensional one
- A method called Singular Value Decomposition for the analysis of co-occurrence patterns is the core

Step of analysis: data reduction

Latent Semantic Indexing

- calculate the best rank k approximation of the key-words document matrix using its SVD

Lower dimensional space of singular vectors
That are called eigen-keywords and eigen-documents

Step of analysis : data reduction

✓ Singular Value Decomposition (SVD)

- “*the best last square appropriated*” in the original frequency matrix for a fixed “k” dimension
- when the “k” value is high, there is an improved approximation to the original matrix. Generally, the “k” value must be big enough to keep its significance for the documents collected, but not so big as to cause a sensation
- values between 10 and several hundred are appropriate as long as the collection of documents is small
- generally, small “k” values (from 2 to 50) are useful for clustering, and larger values (from 30 to 200) are advantageous for forecasting or for classification.

Step of analysis: clustering

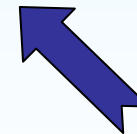
- ✓ cluster analysis on the basis of Euclidean distances computed from one or more quantitative variables

Step of analysis: clustering

Cluster analysis
classifies data
in group whose
characteristics
are a priori
unknown

Hierarchical: each
class belongs to a
wider class

*Non hierarchical: one
partition of the units*



<i>CLUSTER NAME</i>	<i>DESCRIPTIVE TERMS</i>	<i>FREQUENCY</i>	<i>PORTION OF DOCUMENTS</i>
<i>Carige</i>	Access, administration, open, art, insurance, auction, increase, own shares, on-line banking, banker.	193	0.1087936866
<i>Solidity</i>	Stock exchange, economics, events, work, company, finance, Italy, operations, contact, clients.	899	0.5067643743

<i>CLUSTER NAME</i>	<i>DESCRIPTIVE TERMS</i>	<i>FREQUENCY</i>	<i>PORTION OF DOC.</i>
<i>Family</i>	Help, area, tellers, rights, family, internet, logistics, world, new products, personnel.	226	0.1273957159
<i>Outlier 1: Crawling error</i>	Documents, elimination, stylesheet.	70	0.0394588501
<i>Outlier 2: Session over</i>	Attention, connection, error, file, guarantee, inactivity, interrupted, maximized, minute, necessity.	76	0.0428410372
<i>Capital manag.</i>	Centre, credit, market, obligations, reserve, private, financial, finance manage insure	310	0.174746336

	CLUSTER NAME					
NAME OF BANCA	Carige	Solidity	Family	Outlier1: crawling errors	Outlier2: connection errors	Capital manag.
Antonveneta		x	X			
BNL		x				X
Banca della Rete		x				X
Banco di Brescia			X			
Banco di Napoli		x	X		x	X
Banca Lombarda			X			
Banca Intesa	x	x		x		
Bancacis		x				

S. Biffignandi, An analysis of web sites as a communication tool

	CLUSTER NAME					
NAME OF BANCA	Carige	Solidity	Family	Outlier1: crawling errors	Outlier2: connection errors	Capital manag.
Antonveneta		x	X			
BNL		x				X
Banca della Rete		x				X
Bipop		x	X			
Brebanca		x	X			
Carige	x	x	X			X
Carifirenze		x	x			x

S. Biffignandi, An analysis of web sites as a communication tool

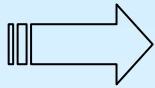
	CLUSTER NAME					
NAME OF BANCA	Carige	Solidity	Family	Outlier1 : crawling errors	Outlier2: connection errors	Capital manag.
Carismi		x	X			
Credem		x			x	X
Carifol		x				
Carispo		x	X			X
Capitalia	x	x				
Farbanca		x				X
Finemiro		x	X			
Fideuram		x				X

S. Biffignandi, An analysis of web sites as a communication tool

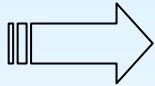
	CLUSTER NAME					
NAME OF BANCA	Carige	Solidity	Family	Outlier 1: crawling errors	Outlier2: connection errors	Capital manag.
IRFIS		X				X
MCC		X				X
Sanpaolo		X	X		X	X
Unicredito		X				
Unicreditbanca		X				
Unicreditimpresa						X
Unicreditprivate		X				
Xelion		X				X

S. Biffignandi, An analysis of web sites as a communication tool

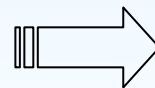
Work in progress



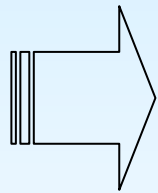
Excluding outlier 1 e 2



Separated analysis of the
home page and of the
associated links



Data set with additional banks



Thank for you attention!!