2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

Identifying ITC Patterns by Industries Using Web Content Mining

Camelia Ratiu-Suciu [1], Florica Luban [1] and Camelia Elena Ciolac [1 +]

[1] Faculty of Management, The Bucharest Academy of Economic Studies, Bucharest, Romania

Abstract. Learning the competitors' IT framework in a knowledge-based economy can become a valuable source of competitive advantage. This paper proposes a framework for knowledge discovery of ITC solution patterns for different industries, by use of web content mining. An application is made in the financial industry, having the companies clustered according to their business coordinates. After having identified the clusters, a quantitative algorithm is provided to assist a financial company make a decision of ITC products acquisition that best suits its business size and that assures technology alignment with industry trends in IT deployment.

Keywords: knowledge, web content mining, ITC patterns, software agents, decision

1. Introduction

In the knowledge-based economy, the competitive environment is often perceived as the primary source of risk in the decision-making process, [1-3]. From the information technology point of view, knowledge about the competitors' ITC capabilities can constitute an important source for competitive advantage for the company, because it can:

- Provide a realistic overview of the competitors' operational capabilities;
- Enlighten the technology trends in the industry in which the company activates, so that it can update its IT infrastructure accordingly;
- Minimize the risks associated with the implementation of new communication technologies;
- Improve the EDI (Electronic Document Interchange) with stakeholders by adapting the company's communication interfaces according to the business partners' IT means of communication.

This paper aims to propose a framework for discovering important knowledge about market players' IT profiles using web content mining, [4].

The valuable knowledge obtained by web content mining will then be used in a quantitative model that we build and will provide an IT solution customized for a specific company whose business coordinates are known.

The model can also provide benchmarking information in terms of information technology solutions' suitability for different industries.

2. Framework for Web Content Mining

The main software providers often offer on their web-sites „success stories" of customers implementing their solutions.

It is in these „success stories" that the web content mining is carried out, because they constitute an underexploited source of knowledge at this moment.

[+] Camelia Elena Ciolac. Tel.: + 4 0727769745 , *E-mail address*: ciolac_c@yahoo.co.uk . This research is sponsored from the European Structured Funds nr. 7832, „PhD Studies and PhD Students in the triangle Education-Research-Inovation, DOC-ECI".

This approach could be criticized in terms of subjectiveness, as it reflects only the successful IT implementations. But, at a deeper level of analysis, this approach also enlightens the migration of a company from an IT solution (a software provider) towards another one, that can be a sign of unsuitability in the relationship IT solution-industry requirements if observed as a mass phenomenon.

The general information offered within a case study, or „success story”, as resulted after a survey of the web-sites cited in bibliography, refers to:

- Quick facts. It is in this section that the customer’s business coordinates are presented: industry, revenue (optional), employees, head quarters/location, key customers (optional).
- Key challenges
- Solution
- Quantitative benefits
- IT framework (optional) including: hardware, database system, server.

This general structure is sometimes enriched with short explanations of the implementation process as well as with quality metrics measured after deployment.

The case study is sometimes accompanied by a CEO’s or CTO’s point of view that summarizes the customer’s perspective upon the benefits brought by the technological change for the business.

In the following paragraphs we shall consider each of the above items and discuss it in detail, along with the possibility of automating the mining process using software agents.

Some issues arise when the web content mining process is to be automated with software information agents:

- (O1) The customers indexing mechanism in the web-site. We have identified various mechanisms of customers indexing, including A-Z names list retrieved at once, A-Z names indexing obtained by parameterized queries for initial letter, unsorted names list.
- From the structural point of view, the customers index can be presented in on unordered list (<ul> and <li> HTML tags) or as titled paragraphs (<h6> HTML tag)
- (O2) Generally, the customers’ index provides direct links to the case study file. On one hand this file can have several formats: html, pdf. On the other hand, sometimes the link is not direct and an intermediate page is displayed in order to emulate the M: N cardinality of the relationship between the set of IT solutions provided and the set of customers.
- (O3) Within the case study file, information is formatted in columns, unordered lists and even tables. The software agent should be capable of delimiting the main areas and not merge information from two distinct areas, even if they are separated by a picture or a horizontal line or different background colors.

As a result, a need for a formal description for the case study’s content arises. A search for short keywords in a large case study text using Boyle Moore algorithm, [5], is a suitable approach

A master agent mines the web in order to retrieve a comprehensive list of software vendors’ websites. The set of URLs is then divided among slave agents.

A slave information agent is attributed a set of URLs of software providers’ websites.

A pseudo code for the slave agent’s action is presented in figure 1.

The procedure new_pattern is responsible of identifying and storing the structure of the newly encountered website. Its strategy is to make a Breadth-First Search in the website’s tree of files in order to get the „success stories” entry point. A file’s children consist of all the files accessible through hyperlinks.

Once identified the entry point in the file tree structure for the „success stories” section, the agent gets the HTML code and identifies which file pattern is used (see O1 – O3 from the previous section).

The result of the parsing process is stored by the agent in a database, in the corresponding fields of the table CASESTUDIES.

The stored information is then subject of a data mining clustering algorithm. The purpose of this methodology is to identify clusters of companies that activate in the same industry and have appropriate business coordinates in terms of number of employees and revenues.

```
for each URL in the set
   retrieve form the database IDr = the ID of the agent
        that web content mined the URL most recently
   if  IDr = NULL then  call  new_pattern
   else
     if self  ID = IDr  then
      try
        locally extract the pattern for the website structure
_1:     apply the pattern to extract list of client companies
        for each company in the list
          use case study pattern to retrieve its IT solution
          store the data in the database
          //in the DB record , the agent specifies its ID
        end for
       catch exception
        // meaning that the known pattern is no longer valid
         call  new_pattern
     else
         query agent IDr to retrieve the pattern to apply
          goto _1
     end if
   end if
end for
```

Figure 1. Pseudocode for a slave agent's actions

The CEO's point of view quoted in some case studies can be subject to Opinion Mining techniques, [6], to discover hidden sentiments of managers with regard to the deployed IT solution that lay beneath the formal quotation.

3. Application on the Financial Services Industry

In order to prove the practical applicability, let us consider the industry „Financial Services". From the case studies presented on the web-sites of the software providers cited in the bibliography section, [7-9], we identified 67 financial companies whose business coordinates could be identified.

For each company, the agents registered the business coordinates, the acquired IT solution, the quantitative benefits obtained as well as the IT framework in which the newly acquired solution was integrated.

Data for a company were integrated from multiple sources, with respect to consistency, as shown in figure 2:

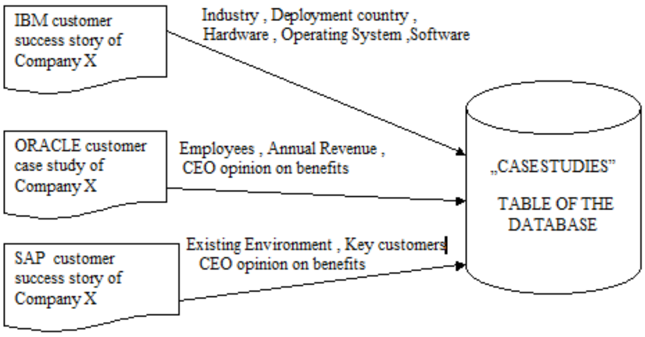
IBM customer success story of Company X
Industry , Deployment country , Hardware , Operating System ,Software
ORACLE customer case study of Company X
Employees , Annual Revenue , CEO opinion on benefits
SAP customer success story of Company X
Existing Environment , Key customers CEO opinion on benefits
„CASE STUDIES" TABLE OF THE DATABASE

Figure 2. Multiple source data integration

A non-hierarchical data mining classification algorithm was considered to be more appropriate for clustering the collected data. Therefore, the k-Means clustering model was chosen in order to group companies with similar business coordinates in clusters.

Running the k-Means on our dataset, considering the variables NREMP (number of employees) and ANNUALREVENUE (the annual revenue of the company) as grouping criteria, with a number of 4 clusters and 20 iterations resulted in the 3D plot presented in figure 3.

Although in figure 3 we plotted company names as well, the following steps of the analysis will be based on company location / deployment country instead of company names, because the research aims to find patterns and not individual behaviours.

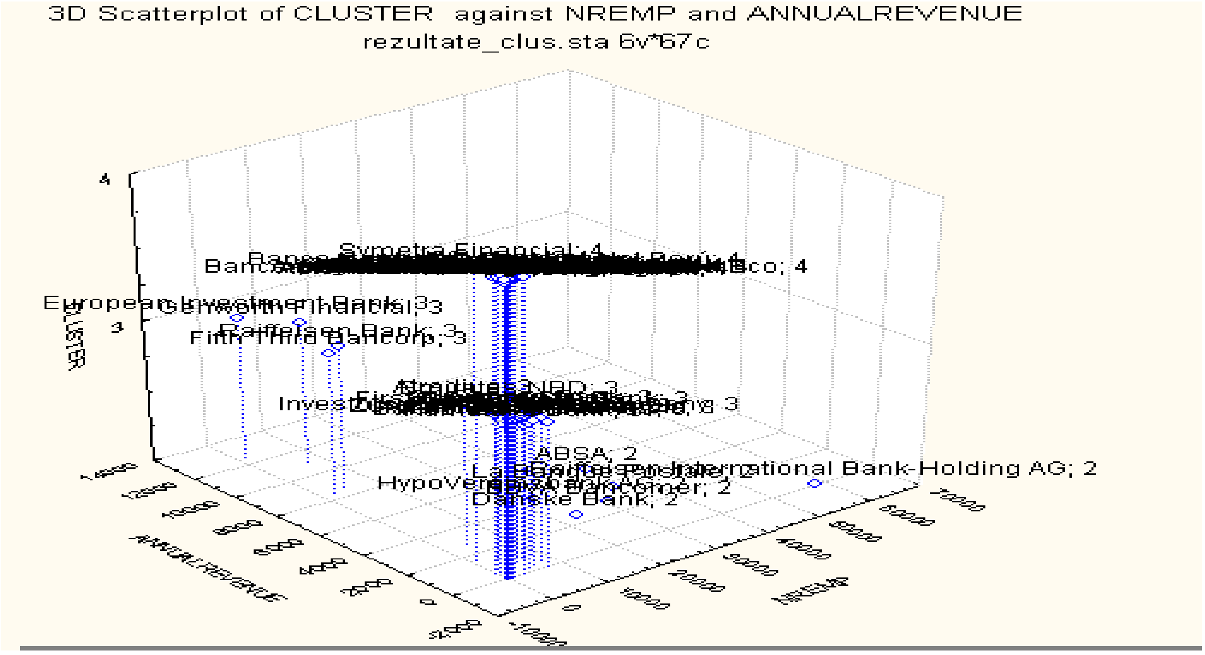

Figure 3. Cluster visual identification

The identified clusters are:

- **Cluster 1**, characterized by a mean in the number of employees of 35000 and a mean in the annual revenues of USD$ 320000 million, contains one company (Table I) with outstanding revenues.

TABLE I. CLUSTER 1 CASES' COUNTRY AND DEPLOYED IT SOLUTIONS

COUNTRY	*DEPLOYED IT SOLUTIONS*
CANADA	ORACLE ONDEMAND, ORACLE FINANCIALS , ORACLE IPROCUREMENT ORACLE PURCHASING

- **Cluster 2** contains 6 companies, (Table II), characterized by a mean in the number of employees of 34564.33 and a mean in the annual revenues of USD$ 3089.5 million.

TABLE II. CLUSTER 2 CASES' COUNTRY AND DEPLOYED IT SOLUTIONS

COUNTRY	*DEPLOYED IT SOLUTIONS*
AUSTRIA	ORACLE DATA INTEGRATOR
DENMARK	PeopleSoft Enterprise Human Capital Manager
France	Oracle Incentive Compensation
GERMANY	Oracle WebLogic Server
MEXICO	Oracle SOA Suite , Oracle BPEL , Oracle University
SOUTH AFRICA	Oracle Balanaced Storecard , Oracle Warehouse Builder

- **Cluster 3** contains 20 companies, (Table III), characterized by a mean in the number of employees of 5620.7 and a mean in the annual revenues of USD$ 2773.884 million.

TABLE III. CLUSTER 3 CASES' COUNTRY AND DEPLOYED IT SOLUTIONS

COUNTRY	*DEPLOYED IT SOLUTIONS*
ARGENTINA	SIEBEL CRM,SIEBEL CONTACT CENTER , SIEBEL SALES
AUSTRALIA	PeopleSoft Enterprise CRM , Oracle Financial Services Applications
AUSTRIA	Oracle Real Applications Clusters, Oracle Discoverer
CHILE	Hyperion Planning , Hyperion Essbase ,Hyperion Web Analysis
INDIA	Oracle Risk Manager
INDIA	Oracle EBusiness Suite , Oracle CRM , Oracle University
LUXEMBOURG	PeopleSoft Enterprise Human Capital Management
MEXICO	Hyperion Planning , Hyperion Essbase , Hyperion Web Analysis
NETHERLANDS	Oracle iLearning
ROMANIA	Oracle iLearning
SWITZERLAND	Siebel Finance, Siebel Marketing
TAIWAN	PeopleSoft Enterprise Human Capital Management
TAIWAN	Oracle Financials, Oracle Risk Manager , Oracle Transfer Pricing
TURKEY	Oracle Financials , Oracle iProcurement , Oracle Assets, Oracle Inventory Manager
UAE	Oracle Human Resources, Oracle Payroll, Oracle Learning Management
UK	Oracle Assets , Oracle Cash Management, Oracle Financials and Sales Analyzer
USA	Hyperion Planning , Hyperion Essbase , Hyperion Web Analysis
USA	Oracle eBusiness Suite On Demand , Oracle iLearning , Oracle Time& Labour Oracle Discoverer
USA	Siebel CRM OnDemand
USA	Oracle OnDemand , Oracle Human Resources ,Oracle iRecruitment , Oracle Learning Management

- **Cluster 4** contains 41 companies, (Table IV), characterized by a mean in the number of employees of 795.3903 and a mean in the annual revenues of USD$ 173.6963 million.

TABLE IV. CLUSTER 4 CASES' COUNTRY AND DEPLOYED IT SOLUTIONS

COUNTRY	*DEPLOYED IT SOLUTIONS*
AUSTRALIA	SIEBEL CRM FINANCIAL SERVICES ,SIEBEL INCENTIVE COMPENSATION MANAGEMENT , ORACLE FINANCIALS
BELGIUM	Oracle Fusion Middleware, Oracle Collaboration Suite, Oracle Portal
BRAZIL	Oracle OLAP , Oracle Warehouse Builder Data Quality
CHINA	DB2 Data Warehouse Edition
CHINA	Oracle Performance Analyzer , Oracle Funds Transfer Pricing , Oracle Enterprise Budgeting and Planning
CHINA	Oracle Discoverer , Oracle Risk Manager , Oracle Assets , Oracle Financials , Oracle Transfer Pricing
COLOMBIA	Oracle Database Enterprise Edition , Oracle Portal
CZECH	Oracle Information Rights Management
CZECH	Oracle Human Resources, Oracle Time and Labour , Oracle Advanced Benefits
DENMARK	Oracle Workflow
EGYPT	Oracle Financial Services Application, Oracle Discoverer
FRANCE	PeopleSoft Enterprise Service Automation , PeopleSoft Performance Management
GERMANY	Siebel Sales
GREECE	Oracle Telesales , Oracle Scripting, Oracle Financials, Oracle Consulting Service
HONDURAS	Oracle Financials, Oracle Fussion Middleware ,Oracle FlexCube
HONK KONG	Oracle Database
HUNGARY	Oracle iLearning
HUNGARY	Oracle iLearning , Oracle Consulting Services
INDIA	Oracle Database
ITALY	Oracle Application Server, Oracle Portal,Oracle WarehouseBuilder , Oracle BPEL
ITALY	Oracle Warehouse Builder , Oracle Workflow
ITALY	Oracle Portal
ITALY	Oracle CRM , Oracle Business Intelligence
ITALY	Hyperion Essbase , Hyperion Web Analysis
MALAYSIA	Oracle Financials , Oracle Fixed Assets ,Oracle Cash Management, Oracle iExpenses
MEXICO	Oracle Discoverer
MEXICO	Oracle Database Standard Edition One
MEXICO	Oracle Business Intelligence , Siebel CRM , Siebel Contact Center , Siebel Marketing
PERU	Oracle Database Standard Edition One
PERU	Oracle Collaboration Suite
ROMANIA	Lotus Domino Collaboration Express
SPAIN	Oracle Financials , Oracle Assets, Oracle Cash Management
SWITZER- LAND	Oracle Database Enterprise Edition
TURKEY	Oracle Data Integrator
UAE	Oracle Teleservice , Oracle Interaction Center
UGANDA	Oracle Financials , Oracle Payroll ,Oracle Purchasing , Oracle Discoverer
UK	Oracle SOA Suite ,Oracle Business Intelligence Enterprise
UK	Oracle Webcenter Suite , Oracle Business Process Manager
USA	Oracle Data Integrator
USA	Oracle CoreId Acess and Identity
VIETNAM	Oracle Data Guard , Oracle Fussion Middleware, Oracle Fiancials , Siebel CRM

Further quantitative modelling is based on the discovered clusters.

Let *Loc* be a vector of deployment countries considered as an indexed set (without duplicates).

Let *Ci* be the set of companies grouped in cluster *i*, *i=1, ..., 4* . For c financial company in Ci let *loc(*c*)* be the deployment country of company c.

In order to use the resulted clusters, suppose that X is a bank that was not considered in the analysis. Bank X is characterized by the following business coordinates:

n = number of employees

r = annual revenue for last year

l = index in the locations vector corresponding to bank X deployment location

Sc = set of key customers of bank X

Sb = set of financial institutions with which bank X develops strong EDI.

The algorithm we propose aims to be a decision assistant tool for problems of IT solution selection and consists of the following steps :

Step 1. Identify the cluster to which bank X belongs to, based on the values of n and r and using an adequate distance metric.

Step 2. Assign bank X to the identified cluster and recompute the cluster mean.

In order to consider possible rearrangements of cases into clusters, a k-Means algorithm iteration should be carried out, with the same parameters as the initial clustering.

Step 3. Let *i* be the cluster where X was assigned to.

Being assigned to cluster *i* , bank X is assimilated with other financial companies that share the same business size, even if situated in the same country or not .

Therefore, bank X should direct its IT investments strategy accordingly and take into account an IT solution similar to its cluster neighbours. Let

PS = { z | z is IT solution and there is at least one company c in Ci that implemented solution z }

PS therefore represents the set of possible IT solutions to be acquired by company X .

Step 4. Compute the scored set of possible IT solutions. Let

PSS = { (z , sz) | z belongs to PS and sz is the score accorded for the option of implementing the software solution z }

Initially,

sz = frequency of use of IT solution z in cluster i

From this first stage, greater scores will be given to frequent solutions that become a „must” for all financial companies characterized by cluster *i* business coordinates.

Step 5. Reshaping scores according to other criteria. Software solutions deployed in the same country as the location of bank X should be given priority. The importance of these IT solutions arises as a consequence of the fact that direct competitors already opted for a specific solution and obtained the same results as bank X, that did not implemented yet that solution (being in the same cluster from the revenues point of view).

This judgement suggests diminishing the scores of software solutions implemented by competitors in cluster *i,* as bank X aims to win a competitive advantage over its competitors that have the same business size. Let

PSS ' = { (z , sz ') | where sz ' = sz – 1 / | PS | if z was deployed by a company c having loc (c) = l }

Step 6. Considering stakeholders' IT solutions.

In order to improve EDI as well as business processes with providers and customers , bank X should consider the stakeholders' acquired IT solutions .

This influence should be of positive sign , increasing the score attributed to a solution that could be provide a common interface with a provider / customer in terms of electronic document interchange. Accordingly, let

PSS '' = { (z , sz '') | where sz '' = sz' + NSz / ($\Sigma_{i=1,4}$ *| Ci) | NSz = the number of stakeholders (Sc union Sb) that deployed the IT solution z }*

The algorithm finally provides a set of software solutions PSS'' ranked against a multitude of factors (both inner – business size, and outer to the company – competitors, competitive advantage, stakeholders).

Step 7. Merging the scored possible solutions set with identified benefits.

In our opinion , the information about the quantitative and qualitative benefits brought by each IT solution as resulted from the web content mining process , should be considered as a differentiating factor after the scoring algorithm reached the sixth stage and provided the set PSS''.

IT solutions' benefits registered in practice depend significantly on business particularities that occur in the internal business processes of the company and cannot be obtained through web content mining. Therefore, they should only be consulted as guidelines and not as the main scoring factor.

The clusters identified and presented in the previous section of this paper, together with this algorithm can constitute a valuable tool in the process of decision-making for acquiring and implementing a new IT solution.

In conclusion, the proposed web content mining framework can be effectively used in the practice of IT management at a microeconomic level and, at the same time, can provide a macroeconomic overview of technical capacity development at industry-level for different countries.

4. References

[1] N. North. *Wissensorientierte Unternehmesfuhrung. Wertschopfung durch Wissen.* , 4. Auflage , Wiesbaden , Gabler Verlag , 2005, ISBN 3834900826

[2] Probst, Raub, Romhardt. *Wissen managen*, Gabler 2006 , ISBN 3834901172

[3] H. Nonaka. *The Knowledge Creating Company* , Oxford : Oxford University Press , 1995, ISBN 0195092694

[4] T. Loton. *Web Content Mining With Java,* England: John Wiley & Sons , 2002

[5] R. S. Boyer, and J.S. Moore. *A fast string searching algorithm.* Communications of the ACM. 20:762-772 , 1977

[6] B. Pang, and L. Lee. *Opinion Mining and Sentiment Analysis* [Electronic Version] , Foundations and Trends in Information Retrieval , Vol. 2, No 1-2 , 2008

[7] IBM Customer Success Stories retrieved January 15 2009 from http://www-01.ibm.com/software/success/cssdb.nsf/ topstoriesFM?OpenForm&Site=corp&cty=en_us

[8] Oracle Customers A-Z retrieved January 15 , 2009 http://www.oracle.com/customers/cust_list_atoz.html

[9] SAP Customer Reference retrieved January 15 , 2009 http://www.sap.com/ecosystem/customers/ customers/ index.epx