17th Symposium IMEKO TC 4, 3rd Symposium IMEKO TC 19 and 15th IWADC Workshop
Instrumentation for the ICT Era
Sept. 8-10, 2010, Kosice, Slovakia

MEASUREMENTS OVER E-COMMERCE ACTIVITIES ON THE WORLD WIDE WEB

ANAGNOSTOPOULOS I.[1], KOUZAS G.[2], LOUMOS V.[3] and KAYAFAS E.[3]

[1]University of the Aegean,
Department of Information and Communication Systems Engineering,
Gorgyras Str., Lymperis premises, Karlovassi, Samos, Greece
e-mail: janag@aegean.gr

[2]University of the Aegean,
Department of Financial and Management Engineering,
Chios, Greece
e-mail: gkouzas@aegean.gr

[3]National Technical University of Athens (NTUA),
Department of Electrical and Computer Engineering,
9, Heroon Polytechneiou Str., Zographou, Athens, Greece
e-mail: {loumos,kayafas}@cs.ntua.gr

Abstract

In nowadays, there is a huge volume of information on the web, which is disseminated to the users in a chaotic way. In order to be accessed, the information must be clustered and classified in appropriate knowledge models. These models will provide and form the initial conditions of a knowledge structure.
A confusing situation exists in the Digital Economy or in the so-called E-Commerce Arena. The way in which economic values and business strategies are created is changing fundamentally and this transforms the general structure of the Economy. A simple survey shows that there is no single, comprehensive and cogent taxonomy of web business models, due to the fact that every classification approach is depending on the importance appointed to the various aspects of the emerging "new economy".
This paper, takes under consideration a web business model taxonomy, proposed by M. Rappa. This taxonomy classifies a web site into a business model according its position in the value chain [1].
Based in the above taxonomy the paper suggests a system, which through information filtering, text retrieval and knowledge techniques, will extract the site content, analyse it and finally classify the e-commerce site. The system uses retrieval techniques, which are emphasised in statistical correlations of words in site content, and scoring content collections with model profiles under the Vector Space Model (VSM) [2],[3].
The anticipated outcome of the proposed system is to identify whether a site is actually offers commerce services through the web. In case of an E-commerce identification, the system proceeds in a business model classification based in Rappa's taxonomy. The system performance concerning the E-Commerce identification and the business model classification was close to 95% and 88% respectively.

1 Introduction

Academic, research, legal and business literature have provided a number of different definitions for electronic commerce, depending on the importance appointed to the various aspects of the emerging "new economy". The most restrictive ones limit electronic commerce to buying and selling goods and services while transferring funds through digital communications. But it is essential that E-Commerce must be observed including all company's internal and external procedures, such as marketing, finance, manufacturing, selling, and negotiation. Rappa proposes a taxonomy, which takes the above under consideration [1]. According to his approach a business model is a method, which allows a business to sustain and promote itself in the market. In other words the business model "spells out" how a company makes money and how it deals with its customers, by specifying where it is positioned in the value chain. The following table (Table 1) presents the Rappa's classification. The type of the web business model is shown in the left column while the corresponding URL examples are presented in the right column respectively. This table provides the reference models as well as the initial learning set for the implemented system, as it will be analytically explained in the following paragraphs.

Table 1 Rappa's web business model taxonomy

Type of model	URL examples
Brokerage model	eTrade, CarsDirect, MetalSite, ChemConnect's World Chemical Exchange, Buzzsaw.com, Accompany, Mobshop, Volumebuy, Etrana, NECX, Yahoo! Store's terms, Yahoo! Stores, ChoiceMall, iMall, Shopping Network, HotDispatch, Amazon's zShops, eBay, AuctionNet, Onsale, Priceline, Respond.com, EWanted, MyGeek.com, DealTime, MySimon, RoboShopper, R U Sure, ShopFind, CareerCentral, BountyQuest
Advertising Model	Excite, AltaVista, Yahoo!, My.Yahoo!, My.Netscape, CyberGold, Netcentives, MyPoints, FreeMerchant, BlueMountain,Buy.com
Infomediary model	NetZero, eMachines.com, Deja.com, ePinions, ClickTheButton, NYTimes.com
Merchant model	Facetime, Amazon, OnSale, Levenger, Gap, LandsEnd, B&N, Eyewire
Manufacturer model	Flowerbud, Intel, Apple
Affiliate model	BeFree, i-revenue.net, AffiliateWorld
Community model	National Public Radio, Deja, ExpertCentral, KnowPost, Xpertsite, Abuzz, Guru, Exp
Subscription model	Slate
Utility model	FatBrain, SoftLock, Authentica

2. Content classification under Information Filtering and Retrieval techniques

The main problem in Information Filtering and Information Retrieval (IF-IR) technology, is to determine in an automatic way what part of information is needed from a large database, such as the world wide web, while minimising the amount of search through irrelevant information. Thus, many heavily visited sites or Portals try to unify the access to multiple information sources, providing by this way intelligent classification of information. Simple query terms can be expanded and interpreted, allowing users to discover relevant information in any requested repository.

Furthermore, when high-speed computers became available for non-numerical work, users thought that a computer would be able to 'read' an entire document collection to extract the relevant documents. It soon became apparent that using the natural language text of a document not only caused input and storage problems, but also left unsolved the intellectual problem of characterising the document content [4]. It is considered that future hardware development may produce more feasible natural language for querying and storage. However, automatic characterisation, in which the software attempts to duplicate the human process of 'reading', is verified as a very sticky problem. More specifically, the 'reading' process involves information extraction, both syntactic and semantic, using it to decide whether each document is relevant or not to a particular query which is dedicated for information classification [5]. The difficulty is not only how to extract the information but also how to use it in order to decide relevance similarity in different clustered knowledge areas. The comparatively slow progress of modern linguistics on the semantic front and the conspicuous failure of machine translation show that these problems are largely unsolved.

Document indexing and categorisation, is a problem which has its origins from the ancient period of human history. The human indexer had to retrieve all the *relevant* documents while at the same had to retrieve as few of the *non-relevant* as possible. In this approach, a sequence of appropriate terms was formed a query mechanism (title, author, etc) that will enable the document to be retrieved in response to that query. Human indexers have traditionally characterised documents in this way when assigning index terms to documents. In the classical problem, the indexer attempts to anticipate the kind of index terms a user would employ to retrieve each document whose content he is about to describe. Implicitly he is constructing queries for which the document is relevant.

This paper is based in the above IF-IR principles and suggests a system, which identifies whether a site has commerce activities and then categorises the E-commerce site in the appropriate model of Rappa's taxonomy. Trying to understand the way the human brain clusters the information, it seems intellectually possible to establish the relevance of a document to a query. However, for such an approach in an intelligent

system, a model has to be constructed, in order to perform knowledge representation and relevance decisions. In the next paragraph follows a description of this model, which is called Vector Space Model.

3. Knowledge representation

In the literature various mathematical models have been proposed to represent information retrieval system and procedures. The Vector Space Model (VSM) represents both queries and documents by term sets and calculates similarities between queries and documents [2]. It is established as an effective and efficient retrieval model and many applications use it to represent knowledge in an IR-IF schema [6]. The VSM assumes that an available term set is used to identify both stored records and information requests.

Both documents and queries can then be represented as term vectors as $D_i = (d_{i1}, d_{i2}, \ldots, d_{in})$ and $Q_j = (q_{j1}, q_{j2}, \ldots, q_{jn})$

Where the coefficients d_{ik} and q_{jk} represent the values of term k in document D_i and query Q_j respectively. In the basic form of VSM, these values are set equal to 1 when term k appears in document D_i or query Q_j, and equal to 0 when the term is absent from the vector. In the present problem, the vector coefficients can take also numeric values depending on the importance of the term in the respective document or query. Documents and Queries are representing site content and web business models respectively.

In this paper, a vector, which holds basic keywords and semantic phrases found in web commerce sites, leads the system's knowledge to the E-commerce Arena. This 'thesaurus vector' is called E-Commerce Vector (ECV) and it is constituted by 539 terms. These terms were extracted from the 75 sites depicted in Table 1.

For the creation of the ECV, thesaurus construction issues and automatic indexing techniques were taken under consideration. In particular, the 'keywords' and 'description' meta names fields of the HTML source code were indexed, and then were further processed through a thesaurus construction technique [7]. This technique uses a stop list to remove common function words from the above meta fields, and a suffix-stripping method to generate word stems for the remaining entries. In order to cover multiword terms and phrases and to increase term retrieval efficiency, every candidate indexing term is set in a normalised form. For the stemming process the Porter's algorithm is used [8].

In this work, site content is identified by a set of terms. As mentioned before the vector coefficients can take also numeric values–weights depending on the importance of the term in the respective document. Weights are assigned to terms as statistical importance indicators. If m distinct terms are assigned for content identification, a site can be conceptually represented as a m-dimensional vector - Site Vector (SV) - as it is shown in equation 1. The used algebraically expression for each term weight is given by equation 2 respectively.

$$SV = (w_1 , w_2 , w_3 , \ldots w_m) \qquad \textbf{(1)}$$

where w_i : is the weight assigned to the i-th term and is 0 for terms not present in site S

$$w_i = tf_i \cdot idf_i = \frac{f_i}{NW} \times \log\left(\frac{n_i}{N}\right) \qquad \textbf{(2)}$$

where: idf_i Inverse document freq.

n_i Number of sites that contain term i in a collection of N sites

f_i Frequency of term i in site S

NW Total number of words in site S

tf_i Normalised frequency of term i in site S

The representation of a Business Model profile under the VSM is similar to a site representation. A profile is the description of a web business model. Each model holds different term weights in order to have a different profile. The advantage of using a common vector for both documents and profiles-queries is that a document can also be used as a query itself. According the above, the Business Model Vector (BMV), which will represent the profile of each web business model is given by equation 3.

$$BMV = (p_1 , p_2 , p_3 , \ldots p_m) \qquad \textbf{(3)}$$

4. The proposed system

Based in the above IF-IR techniques, a system is proposed in order to relatively classify an e-commerce site according the Rappa's reference models. The system will compare the semantic content of the site (pages on the web) with dynamic formatted profiles, which are similarity indicators for business model categorization.

The initial problem is how to define profiles for each e-commerce model. A set of terms must be found, which will characterize each model according its weights. Thus, in order to initialise the business model's profiles, the proposed URLs from Table 1 are inserted for each corresponding web business model. The

weights for each term are calculated by equation 2. As it is obvious, the more the known sites are initially used for each model, the more accurate the profile creation is. So, it is considered that the Brokerage and the Advertising model are better defined through their BMV, in our approach. For this reason Manufacturer, Affiliate, Subscription and Utility models are excluded for this research due to lack of sufficient amount of proposed URL for profile initialisation.

4.1 Profile re-weighting

After site classification, the system will automatically re-weight each term for the corresponding profile vector. This provides the system the ability to survey semantic changes in a site content. The formula for this re-weighting is given by equation 4.

$$rw_{im} = \left(\frac{n_{im}}{N_m} \right) \cdot avg(w_{im}) \qquad \textbf{(4)}$$

where, rw_{im} re-calculated weight of term i in business model m
n_{im} number of documents that contain term i in a collection of N documents in model m
$avg(w_{im})$ the average weight of term i as this is calculated by equation 2 in model m
Then the re-weighted profiles (which stands for the e-commerce models classification terms) are inserted in an inverted file by *pprocx* module. Both *processx* and *pprocx* modules are creating inverted files, which have binary form and are similar to the indexes used in database tables, performing SQL queries [9],[10].

4.2 Scoring web sites and business models

The anticipated outcome of the proposed schema is the relative classification of an E-commerce site to a reference business model. This is performed in the final stage of the proposed system under the *scorex* module.
There are many similarity measure approaches and formulas that can be found in the IF-IR literature. However, the similarity metric used in this paper is the cosine coefficient. This method calculates the angle, which is formed between two vectors, and it is given by taking their scalar product according to equation 5. As this value increases the more similar are the vectors.

$$\text{Similarity } (SV_i, BMV_j) = \frac{|SV_i \cap BMV_j|}{\sqrt{|SV_i|} \cdot \sqrt{|BMV_j|}} = \frac{\sum_K w_{ik}^{SV} \times p_{jk}^{BMV}}{\sqrt{\sum_{k=1}^{t} \left(w_{ik}^{SV}\right)^2} \cdot \sqrt{\sum_{k=1}^{t} \left(p_{jk}^{BMV}\right)^2}} \qquad \textbf{(5)}$$

where w_{ik}^{SV} is the weight of term k in the Site Vector SV_i
p_{jk}^{BMV} is the weight of term k in the Business Model profile Vector BMV_i
t is the total number of the terms (539 in the tested case).
However, scores must only be compared unless they are all on a uniform scale. This can be achieved by normalising the site and profile vectors. As a result we have the following constrains:

$$\left|SV_i\right| = \left|BMV_j\right| = 1 \Rightarrow \left|S(SV_i, BMV_j)\right| \le 1 \ \forall i, j \qquad \textbf{(6)}$$

$$Similarity(SV_i, BMV_j) = \sum_k w_{ik}^{SV} \times w_{jk}^{BMV} \qquad \textbf{(7)}$$

4.3 Relevance Threshold for E-commerce identification and classification

For E-commerce identification and classification, this paper suggest the use of a relevance threshold in the similarity value between the SVs and the BMVs, as explained in the previous paragraph. This threshold is considered to be the minimum value, which will define a web site as an E-Commerce one (EC threshold). In other words, site contents with score above the threshold are consider commercial ones, and those below not.
This threshold in an IF-IR system can be extracted from the tested set which initially feeds the system. In our case 57 sites from the Brokerage, Advertising, Infomediary, Merchant and Community model were used as the tested set. The minimum similarity value, which was taken among the Site Vectors set, compared with all the Business Model Vectors was close to 0.118756. Thus, the threshold value for E-commerce identification in the proposed system is equal to 0.1.
A similar procedure for the definition of the Business models threshold value is taken under consideration. This threshold value is considered as the minimum value, which will classify a web site to its corresponding business model (BM threshold). In other words, site contents with score above this threshold are consider to

belong to the respective model, and those below not. As a testing procedure seems very similar with the previous one, yet more difficulties were presented. As it is shown in Table 1 the set for each tested model isn't the same. Brokerage model in Table 1 hold the most URLs for testing comparing with the other ones. Thus, this model can be considered as the better one for the definition of the BM Threshold. The minimum similarity value, which was assigned among the Brokerage model URLs was close to 0.154864. In the other models this minimum value was presented in range between 0.132383 and 0.171106. Nevertheless, even if the threshold for each profile can be different in this work it is decided to set the BM threshold value equal to 0.15, which seems to be the corresponding threshold for the Brokerage Model.

4.4 System performance

In order to evaluate the system performance in E-commerce identification, 4 test sets were taken under consideration (Table 2). The first set was constituted with 20 known E-commerce sites from the Brokerage, Advertising, Infomediary, Merchant and Community model. Similarly, the second set tested 11 known E-commerce sites from Table 1 (different from the previous set), as well as 5 known non E-Commerce sites. In the third set 14 known E-commerce sites (irrelevant to Rappa's taxonomy), were tested together with 9 non E-Commerce sites. Finally a set with only non E-Commerce sites formed the fourth test set.

Table 2 E-Commerce Identification Performance

E-Commerce Identification Performance						
	E-commerce Sites		Non commerce Sites		Total (%)	Overall (%)
	Successfully Identified	Total	Successfully Identified	Total		
Test1	20	20	-	-	100%	94,87%
Test2	11	11	4	5	93,75%	
Test3	13	14	8	9	91,30%	
Test4	-	-	17	18	94,44%	

As it is shown in Table 2, the system successfully identified all the sites from the first test set, meaning that all the sites were assigned with a score value above the EC Threshold. The performance in this case was excellent, yet this was expected since this was used as an initialization set for the system. Similarly, the system correctly identified all the 11 known E-commerce sites from Test2. However, in this test set a known non E-commerce site was identified as an E-commerce one, and as a result the system performance was rated at 93,75%.

In the third test set the system successfully identified 13 of 14 E-commerce sites and 8 from 9 non E-commerce sites. Finally in fourth test set, which was constituted only with non E-Commerce sites, the system identified 17 from 18 sites. These last test sets (Test3 and Test4) were very interesting because the tested E-commerce sites were not included in the Rappa's web business model classification. However, the system performance in the above cases was in high levels (91.30% and 94.44% respectively).

Table 3 Model Classification Performance

Model Classification Performance				
	E-commerce sites		Total (%)	Overall(%)
	Successfully Classified	Total		
Test1	17	20	85%	87,97%
Test2	10	11	90,91%	

The first two test sets can also be used, in order to measure the system performance in model classification according to the proposed taxonomy, as it is shown in Table 3.

In the first test set, 17 of the 20 E-commerce sites were successfully classified in the appropriate web business model of the proposed taxonomy. In other words, 17 sites were assigned with a score value above the BM Threshold for its corresponding model, while the rest 3 failed to reach it. In the second test set 10 of 11 E-commerce sites were successfully classified. So the performance in model classification ranges between 85 and 91 percent respectively. Despite the fact that the model classification performance is in lower levels, this was expected since in some models (especially in the Infomediary, Merchant and Community) there is a lack of efficient number of proposed URLs, for the definition of the BM Threshold, comparing with the Brokerage and the Advertising model.

As it is obvious, in the third test set the system classified the identified E-Commerce sites, yet this classification cannot be strictly evaluated because the used URLs were not included in the Rappa's taxonomy. Nevertheless, it cannot be ignored that such a classification is fundamentally based in the above taxonomy.

5 Conclusions – Future work

This paper proposes the implementation of a system, which will identify web sites that perform and offer commerce activities and services to the users (E-commerce sites). Furthermore, after the identification the system proceeds to a model classification, based in a taxonomy for web business models firstly proposed by M. Rappa. Sites and models are considered as vectors under the Vector Space Model theory for knowledge representation, and the similarity metric used for comparing these vectors is the cosine coefficient. For the initialisation phase (learning phase) of the system, URLs of the proposed taxonomy are taken under consideration in order to define the knowledge thresholds for the site identification and classification. For the first case the system performance is in high levels (close to 95%). In model classification, the system present lower performance (close to 88%) due to lack of efficient amount of proposed URLs, for the definition of the respective threshold value.

The system model classification performance would be definitely in higher levels, if the proposed taxonomy provide us with more suggested URLs for each model. Four models were excluded in this research, because it was evaluated that will mislead the implemented system. Furthermore, it is considered that the model classification performance will increase if each model is assigned with different Business Model Thresholds. This is an issue, which is left for future work. However, even in such an approach a larger amount of proposed URLs for each model are needed. Future work involves also issues of optimizing the retrieval process, such as relevance feedback and Thesaurus optimization.

6 References

[1] Managing the Digital Enterprise, "Business Models on the Web", URL:<http://ecommerce.ncsu.edu/business_models.html>, 2000.
[2] G. Salton, Automatic Text Processing, Addison-Wesley Publishing Company Inc,1989, 313-326.
[3] G. Salton, A. Wong and C.S. Yang, A Vector Space Model for Automatic Indexing, Communications of the ACM, 18:11, November 1975, 613-620.
[4] C.J. van Rijsbergen, Information Retrieval, Butterworths, London, Second Edition, 1979.
[5] Marie-Francine Moens, Automatic Indexing and Abstracting of Document Texts, Kluwer Academic Publishers, 2000.
[6] V.V. Raghavan and S.K.M. Wong, A Critical Analysis of the Vector Space Model for Information Retrieval, Journal of the American Society for Information Science (JASIS), 37:5, September 1986, 279-287.
[7] G. Salton, Automatic Text Processing, Addison-Wesley Publishing Company Inc,1989, 301-303.
[8] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co. Inc., 1999, Appendix: Porter's Algorithm.
[9] F.W. Lancaster, Information Retrieval Systems: Characteristics, Testing and Evaluation, Second Edition, John Wiley and Sons, New York, 1979.
[10] G. Salton, Automatic Text Processing, Addison-Wesley Publishing Company Inc,1989, 231-240.