BRWM: A relevance feedback mechanism for web page clustering

Ioannis Anagnostopoulos[1], Christos Anagnostopoulos[2], Dimitrios D. Vergados[1] and Ilias Maglogiannis[1]
1 Department of Information and Communication Systems Engineering, University of the Aegean, Karlovassi 83200, Samos – GREECE
2 Department of Cultural Technology and Communication University of the Aegean, Mytiline 81100, Lesvos – GREECE

Abstract. This paper describes an information system, which classifies web pages in specific categories according to a proposed relevance feedback mechanism. The proposed relevance feedback mechanism is called Balanced Relevance Weighting Mechanism – BRWM and uses the proportion of the already relevant categorized information amount for feature classification. Experimental measurements over an e-commerce framework, which describes the fundamental phases of web commercial transactions verified the robustness of using the mechanism on real data. Except from revealing the accomplished sequences in a web commerce transaction, the system can be used as an assistant and consultation tool for classification purposes. In addition, BRWM was compared with a similar relevance feedback mechanism from the literature over the established corpus of Reuters-21578 text categorization test collection, presenting promising results.

1 Introduction

This paper presents an algorithm, which clusters web pages that offer commercial services, according to an e-commerce framework and under the basic concepts from the field of information retrieval. The paper is organized as follows. The next section presents an overview of several e-commerce/business frameworks, while it analyses the model is used for the purposes of this work. Section 3 describes some basic information retrieval procedures and mechanisms made in order to conceptually represent the web transactions and phases. The next section presents the main parts and procedures of an information system, which identifies commercial services on the web according to the followed e-commerce framework

Please use the following format when citing this chapter:
Anagnostopoulos, Ioannis, Anagnostopoulos, Christos, Vergados, Dimitrios, Maglogiannis, Ilias, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 44–52

and a proposed relevance feedback mechanism. In parallel, the algorithm is evaluated over an established text categorization test collection from the information retrieval literature and is compared with a similar relevance feedback algorithm. Finally, the paper ends with the results derived over the followed e-commerce framework as well as with a discussion over the potential applications of the presented work.

Table 1. Transaction phases and types of e-commerce pages according to the BMF

Transaction Phase - PT	e-commerce page type	web pages (per type)	web pages (per PT)
Knowledge	Query engines homepages	322	1305
	Directory pages	339	
	Product information pages	365	
	Registration pages	279	
Intention	Product catalogue pages	372	777
	Order – Payment pages	405	
Contracting	Terms and conditions pages	387	387
Settlement	Settlement monitoring pages	313	1355
	Digital delivery pages	364	
	Contact and complaint forms	358	
	After sales support pages	320	

2 The web transactions framework

Numerous different frameworks for the analysis of web commerce/business models have already been proposed in the literature. These frameworks are generally using different approaches to identify, classify and analyse commercial activities [1], [2], [3], [4], [5], [6], [7]. Among the above-mentioned models, the Business Media Framework (BMF) was used for the purposes of our work [1]. This taxonomy relates and interprets the components of the general media model in the business context, thus providing a means for the design and management of business communities. The framework distinguishes four distinct views and four transaction phases of a web commercial transaction. The community view deals with the aspects relevant for modelling the community, such as its organisational structure and shared roles, the protocols, the interests and values behind it, as well as its languages. The implementation view describes the specified community design, such as the community view specifications, the data structures and the business processes on the services offered by the service layer. The transaction view provides the generic interaction or communication services such as the signalling intentions, the contracting and agreement on contracts and or the performance of the transaction in the settlement of contracts. Finally, the infrastructure view provides communication and coordination mechanisms for relating agents of the transaction view as well as the agents taking the roles modelled in the community view, as defined in the implementation view [1].

In parallel BMF also distinguishes four transaction phases. In the knowledge phase a common logical space between agents is being established and information about the transaction is gathered and processed. The intention phase includes services for analysing and activating information acquired in the knowledge phase. In the contracting phase a "contract" is being negotiated between agents while the settlement phase refers to the settlement of the "contract" aforementioned, which actually means the realisation of the web transaction. The distinction of these four phases identifies the structural changes that electronic commerce has brought to traditional commerce methods. Table 1 presents the four phases and the amount of the collected web pages, which were used as the training material. The total sample set consists of 3824 e-commerce pages of several extension formats. These web pages were collected and validated by experts according to BMF. As a result, each web page depicted in Table 1 corresponds to one e-commerce type and one transaction phase. However, a respective data sample that consist of 2134 web pages and do not describe commercial transactions (web pages irrelative to web commerce transactions), was collected automatically using a meta-search engine tool. This tool collects randomly web pages from specified search engine directories and its functions are described in [8].

3 Feature selection

This section describes the feature extraction procedure for the training sample. The training sample consists of twelve classes from which, eleven of them correspond to the BMF transaction phases (knowledge, intention, contracting, settlement) and one class correspond to web pages that do not offer commercial services. Common information filtering techniques such as stop lists, character filters and suffix-stripping methods were initially used for reducing the large amount of the indexed terms. The Information Gain (IG) technique was adopted for feature selection. This technique measures the statistical dependencies between a term and the categorised class based on the entropy. Thus, terms with small information gain are discarded [9]. Rather than evaluating the entropy of a term distribution among a set of documents as is done for the signal-to-noise ratio technique, in the specific technique the entropy of the class distribution is taken under consideration. More specifically, let C denote a random variable used for observing the k possible class labels for the training documents.

$$Entropy(C) = -\sum_{i=1}^{k} P(c_i) \log P(c_i) \quad \textbf{(1)}$$

In IG, entropy measures the homogeneity of the training set D^* with respect to the class distribution which governs C according to Equation 1, where $P(c_i) = (n_{c_i}/n)$, denotes the probability of observing a training document with category c_i and $0log0$ is defined to be zero for all entropy evaluations. Based on this interpretation of entropy, the discriminative power of a particular index term can be measured as follows. Let t and $\bar{t}$ denote the presence and absence of term t, respectively, and T be a binary random variable taking on the values t and $\bar{t}$. The conditional entropy of the random class variable C given T is defined according to

Equation 2, where $P(t)=(n(t)/n)$ and $P(\bar{t})=(n(\bar{t})/n)$ denote the proportions of training documents in which term t is present and absent, respectively. The conditional probabilities are estimated by $P(c_i|t)=(n_{c_i}(t)/n(t))$ and $P(c_i|\bar{t})=(\bar{n}_{c_i}(t)/\bar{n}(t))$. Finally, the Information Gain of term t is defined from Equation 3, as the expected reduction in entropy caused by partitioning the set of training examples D according to the presence or absence of term t. However, by using elementary probability calculus, Equation 3 can be transformed to Equation 4, where the probabilities that a class c_i and a term t do or do not co-occur can be derived from the probabilities introduced as $P(c_i,t)=(n_{c_i}(t)/n)$ and $P(c_i,\bar{t})=(n_{c_i}(\bar{t})/n)$ respectively.

$$\begin{aligned}&Entropy(C\,|\,T) = P(t)Entropy(C\,|\,t) + P(\bar{t})Entropy(C\,|\,\bar{t}) = \\ &= -P(t)\sum_{i=1}^{k}P(c_i\,|\,t)\log P(c_i\,|\,t) - P(\bar{t})\sum_{i=1}^{k}P(c_i\,|\,\bar{t})\log P(c_i\,|\,\bar{t})\end{aligned} \quad (2)$$

$$Gain(t) = Entropy(C) - Entropy(C\,|\,T) \quad (3)$$

$$Gain(t) = \sum_{i=1}^{k}P(c_i,t)\log\frac{P(c_i,t)}{P(c_i)P(t)} + \sum_{i=1}^{k}P(c_i,\bar{t})\log\frac{P(c_i,\bar{t})}{P(c_i)P(\bar{t})} \quad (4)$$

Using the IG technique, 1063 terms were finally selected in order to compose the vector that represents web pages (Web Page Vector – WPV). The WPV characterize a web page by assigning a unique profile of weight values that depend on the importance of each term in the tested web page. In other words, weights are assigned to terms as statistical importance indicators. If m distinct terms are assigned for content identification, a web page is conceptually represented as an m-dimensional vector, named WPV. Equation 5, highlights the lnc formula used for the weighting mechanism, which is based on the SMART system as described in [10]. Web Page Vector is defined as $WPV_i = \{w_{i1}, w_{i1}, \ldots, w_{ik}\}$, while the weight of term k in the i^{th} web page is normalized using the cosine length of the vector, where l equals to 1063 and corresponds to the total amount of the used terms.

$$w_{ik}^{SV} = (\log(tf_{ik}) + 1)\cdot(\sum_{k=1}^{1063}[\log(tf_{ik}) + 1]^2)^{-1/2} \quad (5)$$

4 System architecture

Based on the above information filtering techniques, the proposed system relatively classifies web pages in twelve categories. It compares the content of the web pages with dynamic profiles, which are similarity indicators for the twelve categories (type of web pages). In order to relatively classify web pages, the proposed system uses similarity threshold values between the *WPVs* and the descriptor vectors. Each threshold is a minimum value dedicated to assign a category label to a tested web page. Thus, web pages with scores above these thresholds are considered to belong to the respective categories, while those with lower scoring values do not. In order to define the respective threshold values we used half of the validated web pages and the PCut thresholding strategy [11], [12].

According to this method for each category c_j, the method sorts the tests documents by score assigning a positive decision to each of the k_j top-ranking documents, where $k_j = P(c_j) \times x \times m$ is the number of documents assigned to c_j and $P(c_j)$ is the prior probability for an arbitrary document that is member of category c_j. The PCut strategy is parameterized by x (fine-tuning parameter), which reflects the average number of documents where the system assigns to a category, m is the number of categories, n represents the number of documents in the validation set and assuming that one scoring value is produced by the classifier for each web page-category pair.

4.1 The proposed relevance feedback mechanism

The scoring mechanism is performed by calculating the scalar product according to Equation 6, where the more this value increases, the more similar the vectors are. The denominator normalizes the similarity comparisons between WPV_i and D_j where w_{ik}^{WPV} is the weight of term k in WPV_i, and p_{jk}^{D} corresponds to the ltc weighting scheme of term k in each descriptor vector D_j as defined in Equation 7. This weighting scheme uses cosine normalization of logarithmic term frequency by the inverse document frequency.

$$(WPV_i,D_j)=\frac{|WPV_i \cap D_j|}{\sqrt{|WPV_i|}\cdot\sqrt{|D_j|}}=\sum_k w_{ik}^{WPV}\times p_{jk}^{D} \quad (6)$$

$$p_{jk}^{D}=(log(tf_{jk})+1)\cdot log(N/n_k)\cdot(\sum_{k=1}^{1063}[(log(tf_{jk})+1)\cdot log(N/n_l)]^2)^{-1/2} \quad (7)$$

After web page classification, the system automatically re-weights the terms of each descriptor vector. The re-weighting is calculated according to a proposed modification of the Rocchio's type for relevance feedback, called Balanced Relevance Weighting Mechanism - BRWM and is defined from Equation 8.

$$D_{j_{new}}=D_{j_{old}}+c\left(1+\frac{n_{rel}}{n_{rel}+n_{irr}}\right)\sum_{r=1}^{n_{rel}}\frac{WPV_r}{n_{rel}}-c'\left(1-\frac{n_{rel}}{n_{rel}+n_{irr}}\right)\sum_{n=1}^{n_{irr}}\frac{WPV_n}{n_{irr}} \quad (8)$$

In the above equation, D_{new} and D_{old} are the re-calculated descriptor vector and the initial descriptor vector, n_{rel} and n_{irr}, stands for the amount of the already recognized relevant and irrelevant pages in respect to a specific category, WPV_r and WPV_n are the relevant and the irrelevant web page vectors, while c *and* c' are fine tuning constants. BRWM re-weights query terms by adding the weights from the actual occurrence of those query terms in the relevant web pages, and subtracting the weights of those terms occurring in the irrelevant web pages. The contribution of the web pages that are not related to a specific information area is to modify the weighting of the terms coming from relevant web pages.

4.2 Evaluation of BRWM

Before the evaluation of the system with real data, we measured the accuracy of BRWM over an established text collection and we compared the results with a similar relevance feedback algorithm (Findsim), which was tested over the same text collection [13], [14]. In particular, we used the corpus of Distribution 1.0 of

Reuters-21578 text categorization test collection. This collection consists of 21578 documents selected from Reuters newswire stories. The documents of this collection are divided into training and test sets. Each document has five category tags, namely, *EXCHANGES, ORGS, PEOPLE, PLACES,* and *TOPICS.* Each category consists of a number of topics that are used for document assignment. This evaluation is restricted to the *TOPICS* category. In particular, we used the Modified Apte split of Reuters-21578 corpus that consist of 9603 training documents, 3299 test documents and 8676 unused documents. The training set was reduced to 7775 documents as a result of screening out training documents with empty value *of TOPICS* category. There are 135 topics in the *TOPICS* category, with 118 of these topics occurring at least once in the training and test documents. The experiment took place with all of these 118 topics despite the fact that three topic categories with no occurrence of training set automatically degrade the performance of the system.

In contrast to information retrieval systems, in text categorization systems a retrieval output is not appeared. Instead, a number of topics occur and for each topic the document collection is partitioned into training and test cases. The training set contains only positive examples of a topic. In this sense, the training set is not a counterpart of the retrieval output due to the fact that there are not any negative examples. However, a training set for a topic that consists of positive and negative examples can be constructed, under the assumption that any document considered as positive example for the other topics and not in the set of positive examples of the topic at hand is a candidate for being a negative example of this topic. Table 2 presents the ten most frequent topics in the category TOPICS of the Reuters collection as well as the respective amounts of the training and testing sets. The maximum number of positive examples per topic in the corpus is 2877 and the average is 84. The size and especially the quality of the training set is an important issue in generating an induction rule set. In the experiment that took place the training set for each topic consist of all the positive samples, while the negative samples were selected from other topics. The size of the selected negative samples was fixed at the 50% of the positives examples. Finally, the Information Gain technique was used since this was the feature selection mechanism used in the compared studies of [13], [14], [15]. According to these papers it was concluded that that the precision or the accuracy of the rules with the Information Gain metric was 3% better than that of rules with CHI metric χ^2. Table 2 also depicts the experimental results deriving from the comparison between BRWM and Findsim over the Reuters-21578 corpus. The comparison is made over the first ten topics and over all the topics of the collection (in average values) and the results are measured in terms of the breakeven point in the precision-recall diagrams of each topic. Precision is defined as the fraction of retrieved web pages, which are relevant to a specific category, while recall is the fraction of relevant web pages, which have been retrieved in respect to the twelve categories. The break-even point is defined as the point where precision is equal to recall.

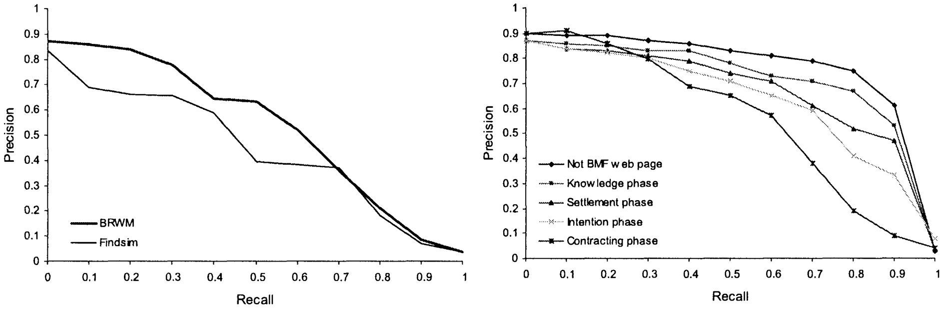

Fig. 1 . Precision-Recall diagrams over: (a) Reuters-21578 corpus, topic "Money-fx", category TOPICS, (b) BMF test sample (average values)

Table 2. Training/Testing sets for the category TOPICS and compared results (breakeven point values) over the Reuters-21578 corpus

Training set	Testing set	Topic of category TOPICS	Findsim	BRWM
2877	1087	Earn	92.9%	90.2%
1650	719	Acq	64.7%	80.6%
538	179	Money-fx	46.7%	56.7%
433	149	Grain	67.5%	63.5%
389	189	Crude	70.1%	68.3%
369	118	Trade	65.1%	60.3%
347	131	Interest	63.4%	67.8%
197	89	Wheat	68.9%	64.7%
212	71	Ship	49.2%	54.1%
182	56	Corn	48.2%	55.3%
		Average top 10	63.7%	66.2%
		Average all	61.7%	64.8%

Topics Earn and Acq were distinguished better among the rest ones on the tested corpus, especially for BRWM. In particular, the breakeven point for the topic Earn was measured above 90% for both algorithms, while for the rest first ten topics the breakeven points presented some fluctuations. However, these fluctuations were expected since the amounts of the training sets are not equal according the Modified Apte split of the collection.

Breakeven point values (in percent) are computed on top 10 topics and on overall 118 topics. In particular, the breakeven point values over the first ten topics were measured equal to 66.2% for BRWM and 63.7% for Findsim, while for the whole collection equal to 64.8% and 61.7% respectively, as presented in Table 2. The results indicated that BRWM presented a better performance in relation to Findsim. Figure 1a presents the precision-recall diagrams of the two compared algorithms in topic "Money-fx" for the category TOPICS. In the precision-recall diagram, labels a and b highlight the breakeven point values that correspond to BRWM and Findsim.

Table 3. Results over the BMF data set

web page type	Break-even point	web page type	Break-even point
Query engines homepages	77.6%	Terms and conditions pages	58.8%
Directory pages	72.8%	Settlement monitoring pages	81.3%
Product information pages	71.2%	Digital delivery pages	77.9%
Registration pages	76.9%	Contact and complaint forms	78.6%
Product catalogue pages	62.6%	After sales support pages	80.4%
Order – Payment pages	68.8%	Not BMF web pages	81.8%

5 Results and discussion

This section presents the results derived from the BRWM algorithm over the rest half amount of the validated BMF data set of Table 1. In particular, Table 3 presents the break-even point values for the eleven type of e-commerce web pages that correspond to the four transaction phases as well as for the web pages that do not offer commercial services (totally twelve categories). According to these values, Figure 1b presents the average precision-recall diagrams for the web pages that correspond to the knowledge, intention, contracting and settlement BMF transaction phase as well as to the page that do not correspond to either of these phases.

The breakeven point for the not BMF related web pages was measured at 81.8%, while for the BMF web pages the average break-even points were measured at 74.6%, 79.6%, 65.7% and 58.8% (knowledge, settlement, intention and contracting phase respectively). Similarly with the tests made over the Reuters-21578 corpus, these variations were expected due to the fact that the amounts of the training sets are not equal among the respective transactions phases of the BMF.

Under the fact that each web page corresponds to one e-commerce type and one transaction phase of the Business Media Framework and this framework analyse an e-commerce model into a series of concurrent sequences, the proposed web information system algorithm can be used in order to identify and classify commercial services and transactions on the web. However, except for classification purposes, the system can be exploited for quantifying e-commerce ontologies and roles. In other words, the system can be either used locally in commercial servers for monitoring customer behaviour directly through local information, or it can be launched independently to a portal, in order to survey and measure commercial activities, services and transactions on the web.

References

1. Klose M., Lechner U., 'Design of Business Media - An integrated Model of Electronic Commerce', In: Haseman, W.D.; Nazareth, D.L. (eds.), Proceedings of the Fifth Americas Conference on Information Systems (AMCIS'99), pp. 115-117, Milwaukee, WI, August 13-15, 1999.
2. Mahadevan B., 'Business Models for Internet-Based ECommerce: An Anatomy', California Management Review, Vol.42, No.4, 2000.

3. Timmers P., '*Business Models for Electronic Markets*', In: Gadient Y., Schmid B. F., Selz D., EM - Electronic Commerce in Europe, EM - Electronic Markets, Vol. 8, No. 2, July 1998.
4. Lawrence E., Corbitt B., Tidwell A., Fisher J., Lawrence J., 'Internet Commerce Digital Models for Business', John Wiley & Sons, Brisbane, 1998.
5. Selz S., 'Web Assessment: A model for the Evaluation and the Assessment of Successful Electronic Commerce Applications', International Journal of Electronic Markets 7(3).
6. Schmid B.F., Lindemann, M.A, 'Elements of a reference model for electronic markets', Proceedings of the Thirty-First Hawaii International Conference on System Sciences, vol.4, pp. 193-201, 1998.
7. Schmid, B., 'What is new about the Digital Economy', Electronic Markets, vol.11, no.1, 04/2001.
8. Anagnostopoulos I., Psoroulas I., Loumos V. and Kayafas E., Implementing a customised meta-search interface for user query personalisation, IEEE 24th International Conference on Information Technology Interfaces, ITI 2002 pp. 79-84, June 24-27, 2002, Cavtat/Dubrovnik, CROATIA.
9. Yang Y and Pedersen J (1997) A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference in Machine Learning, ICML'97, pp. 412 – 420, , 1997, Nashville, TN, USA.
10. Buckley C, Salton G and Allan J (1993) Automatic retrieval with locality information using SMART. In: Proceedings of the 1st Text REtrieval Conference (TREC-1), pp. 59-72, 1993, Gaithersburg, MD, USA.
11. Lewis D., An evaluation of phrasal and clustered representations on a text categorisation task, 15th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 92), pp.37-50, 1992.
12. Yang Y., An evaluation of statistical approaches to text categorization, Journal of Information Retrieval, 1(1/2), pp.67-88, 1999.
13. Dumais S, Platt J, Heckerman D and Sahami M, Inductive learning algorithms and representations for text categorization. In: Proceedings of the 7th international conference on Information and knowledge management, ACM Press 1998, Location, pp. 148-155.
14. Alsaffar A, Deogun J and Sever H, Optimal queries in information filtering. Lecture Notes in Artificial Intelligence (LNCS Series), 1932:435-443.
15. Sever H, Gogur A and Tolun M., Text Categorization with ILA. Lecture Notes in Computer Science – LNCS, 2869:300-307.