Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 91 (2016) 131 – 140

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2016)

Multi-criteria web mining with DRSA

Couto, Ayrton Benedito Gaia do[a*], Gomes, Luiz Flavio Autran Monteiro[b]

[a]System Analyst, Brazilian Development Bank (BNDES), Av. República do Chile, 100, Rio de Janeiro-RJ, 20031-917, Brazil
[b]Professor, Ibmec/RJ, Av. Presidente Wilson, 118, 11th floor, Rio de Janeiro-RJ, 20030-020, Brazil, autran@ibmecrj.br

Abstract

This study demonstrates the application of the Dominance principle to a particular case of web (World Wide Web) content search under Multi-criteria approach: searching for "Rio de Janeiro" (City and/or State, in Brazil) followed by other attributes (or criteria). It is known that depending on the content of research that is carried out through a "seeker" ("search engine") on the Internet, the result may fall short of the desirable, in terms of quantity and quality of the sites returned. The Dominance principle, subsequent to treatment of the collected information (unstructured data) on the Internet, aimed at revealing patterns (or logical rules) on a set of information and showed how a web content search can become more effective at a significant universe of information. Other techniques and tools have been applied to mining content on the Web, and as shown in this study. The choice of the Dominance principle associated to Rough Set Theory as Multi-criteria decision technique is due to the possibility of inaccurate data (inconsistent) and the need for treatment of these inaccuracies when processing an information system (data table) under a mathematical perspective, and do not need a history of these data. The use of Rough Set Theory and the Dominance principle associated with the probabilistic relationship between conditions and decisions in decision algorithms, is showed by the possibility of there being uncertain data to yield an essential set of effectively consistent information.

Peer-review under responsibility of the Organizing Committee of ITQM 2016

Keywords: Dominance principle, Rough set theory, Multi-criteria analysis, Web mining ;

1. Introduction

The majority of users realize the Internet information extraction from search engines or Web browsers. These search engines do not necessarily return the information users want, both in terms of volume and in terms of

* Corresponding author. Tel.: +55 21 2172-7658.
E-mail address: ayrtoncouto@gmail.com.

Peer-review under responsibility of the Organizing Committee of ITQM 2016
doi:10.1016/j.procs.2016.07.050

content. The concept of "web mining" or "data mining of Web" can be defined as the process of discovery and analysis of useful information from the data originated. Includes three types of information: data in Internet; data "log" of Internet access servers, user registration, profiles, etc.; and web structure data. In the case of web mining content, the goal is to identify "patterns" of behavior and extract knowledge from a set of data related to documents (text, image, audio, video, etc.) stored in tables within a web environment. In the case of web data, unstructured documents with different attributes which may have similar semantics in the context of web information. The knowledge discovery "hidden" on the Internet is one of the major features of the process "web mining". This discovered knowledge can be very useful to decision makers, helping them to identify abnormal or unknown behavior in the use or the content of the Internet [1]. Currently, the use of the term "science of data" is increasingly common, as well as the term "big data." Here "science of data" is the study of the data knowledge extraction (heterogeneous and unstructured - texts, images and videos from networks with complex relationships between its entities). As examples, Paypal and Google use predictive models to supported business on the Internet [2].

In the context of this study, we used Google to search the set of URLs (Universal Resource Locator) and corresponding sites summaries with one or more words, particularly about the City and/or State "Rio de Janeiro" (Brazil) followed by other attributes. Depending on the research that takes place, the result can be a significant amount of URLs arranged under a "ranking" ("PageRank", in the case of Google). This "ranking" indicates the most searched sites (quantity and quality) in a descending order according seeker's own criteria [3], [4]. For the result of this search was as effective as possible, this study was guided then, the following research question: "How to identify patterns (or rules) in the Web mining under Multi-criteria approach?". The choice of the Rough Set Theory (RST) and the Dominance principle (Dominance-based Rough Set Approach, DRSA) as tools to support Multi-criteria decision justified by the possibility of inaccurate data (inconsistent) and the need for treatment of these inaccuracies; and the ability to process an information system (or data table) in a mathematical perspective as well, do not need a data history as required by Fuzzy Sets (Fuzzy Set Theory) proposed by Lotfi Asker Zadeh in 1965 [5]. As support for Multi-criteria analysis, we used the jMAF software (Dominance-based Rough Set Data Analysis Framework) [6], given for purposes of research at the Computer Science Institute, Poznan University of Technology, Poland. This study includes a brief approach on Rough Set Theory (RST) and the Dominance principle (DRSA) - Sections 2 and 3, respectively; the application of the Dominance principle to a specific case, Section 4; and ends with the conclusions and directions for future studies, Section 5.

2. Rough Set Theory

RST had its origin with Zdzislaw Pawlak: it proposes the treatment of data uncertainty using “lower and upper approximations” for a data set [8]. One of its concepts, the “indiscernibility relation,” identifies objects that have the same properties, i.e., “indiscernible” objects, to be treated as similar or identical. An information system can be defined as a tuple S = (U, Q, V, f), where U is a finite set of objects, Q is a finite set of attributes, V= $\bigcup_{q \in Q}$ Vq, where Vq is the domain of attribute q, and f: U χ Q → V is a total function such that f(x, q) ∈ Vq for every q∈Q, x∈U, known as an “information function” [8]. Given an information system S = (U, Q, V, f), P ⊆ Q, and x,y ∈ U, we say x and y are “indiscernible” through the set of attributes P in S if f(x,q) = f(y,q) for all q∈P. Therefore, all P ⊆ Q generate a binary relation in U, known as an “indiscernibility relation”, denoted by IND(P). Given that P ⊆ Q and Y ⊆ U, the lower ($\underline{P}$Y) and upper approximations ($\overline{P}$Y) are defined as:

$$\underline{P}Y = \cup\{X \in U/P : X \subseteq Y\}; \quad \overline{P}Y = \cup\{X \in U/P : X \cap Y \neq \emptyset\} \tag{1}$$

The difference between $\underline{P}$Y and $\overline{P}$Y is called the “boundary region” of Y:

$$BNP(Y) = \underline{P}Y - \overline{P}Y \tag{2}$$

There is also the concept of accuracy:

$$\alpha P(Y) = \text{card } \underline{P} / \text{card } \overline{P} \tag{3}$$

which captures the degree to which the knowledge of set Y is complete. There are two more fundamental concepts in RST: an information system's “reduct” and “core”. The reduct is its essential part, i.e., the subset of attributes that provides the same quality of classification as the original set of attributes (it allows one to make the same decisions as if all condition attributes were there). The core is the most important subset of this knowledge; CORE(**P**) = ∩ RED(**P**), where RED(**P**) is the family of all "reducts" of **P** [8], [9].

A "decision rule" is an expression of the form "*if ... then ...*" or $\Phi \rightarrow \Psi$ where Φ and Ψ represent the condition and decision, respectively, of the decision rule. Thus, a decision rule $\Phi \rightarrow \Psi$ is "admissible" in a set S if $| \Phi |_S$ is the union of elementary-C sets (condition), if $| \Psi |_S$ is the union of elementary-D sets (decision) and $| \Phi \wedge \Psi |_S \neq 0$. An example with six stores and four attributes [7] – Table 1:

Table 1. Example with six stores and four initial attributes

Store	E	Q	L	P
1	High	Good	No	Profit
2	Average	Good	No	Loss
3	Average	Good	No	Profit
4	None	Average	No	Loss
5	Average	Average	Yes	Loss
6	High	Average	Yes	Profit

And the corresponding decision rules:

(E, average) and (Q, good) → (P, loss)
(E, none) → (P, loss)
(E, average) and (Q, average) → (P, loss)

3. Dominance principle

The key aspect of a Multi-Criteria decision is considering objects that are described by multiple criteria and that represent conflicting points of view. Criteria are attributes in domains with an ordering preference; e.g., in choosing a car, one may consider the price and fuel consumption to be characteristics that should serve as criteria in its acquisition, as one usually considers a low price to be better than a high price and moderate fuel consumption to be more desirable than high consumption. In general, other attributes such as colour and country of origin, the domains of which have no ordering preference, are not considered to be decision criteria – they are regular attributes. Therefore, the RST approach does not allow one to analyse Multi-criteria decision problems because the analysis uses only regular attributes. Moreover, one cannot identify inconsistencies that violate the following Dominance principle: "objects with a better evaluation or having at least the same evaluation (decision class) cannot be associated to a worse decision class, all decision criteria being considered". RST ignores not only the preference ordering in the set of attributes' values but also the "monotonic" relation of objects' evaluations regarding the condition attributes' values and decision attributes' values' order of preference (classification or degree of preference) [10], [11]. This problem is treated in an extension of RST called Dominance-based Rough Set Approach or DRSA [10], in which indiscernibility relations are replaced with dominance relations in the approximations of decision classes. Furthermore, due to the preferential ordering between decision classes, sets become approximations known as unions of "upward" and "downward" decision classes. Thus, for a tuple S = (U, Q, V, f), set Q is generally divided into condition attributes (set C) and decision attributes (set D). Assuming all condition attributes ($q \in C$) are decision criteria, S_q represents a non-classifiable relation in U with respect to criterion q such that xS_qy denotes "x is at least as good as y in regards to criterion q". Assuming the set of decision attributes D defines a partition of U into a finite number of classes, $Cl = \{Cl_t, t \in T\}$, $T = \{1, ..., n\}$ is a set of these classes such that each $x \in U$ belongs to one and only one $Cl_t \in Cl$. These classes are assumed to be ordered, i.e., for every $r,s \in T$ such that $r > s$, objects of Cl_r are preferable to objects of Cl_s. Therefore, objects can be approximated by unions of "upward" and "downward" decision classes, respectively: $Cl_t^{\geq} = \bigcup_{s\geq t} Cl_s, Cl_t^{\leq} = \bigcup_{s\leq t} Cl_s$, t=1, ...,n. The indiscernibility relation is thus substituted with a dominance relation. One says that x dominates y regarding P ⊆ C, denoted xD_Py, if xS_qy for all q∈ P. The dominance relation is reflexive and transitive. Given that P ⊆ C and x ∈ ∪, the "granules of knowledge" used in the DRSA approximations are:

- a set of dominating objects x, called the P-dominating set: $D_P^+(x) = \{y \in \cup: yD_Px\}$,
- a set of objects dominated by x, called the P-dominated set: $D_P^-(x) = \{x \in \cup: xD_Py\}$.

Using the $D_P^+(x)$ sets, the P-lower and P-upper approximations of $Cl_t^{\geq}$ are: $\underline{P}(Cl_t^{\geq}) = \{x \in \cup: D_P^+(x) \subseteq Cl_t^{\geq}\}$, $\overline{P}(Cl_t^{\geq}) = \bigcup_{x\in Cl_t^{\geq}} D_P^+(x)$, for *t*=1,...,n. Analogously, the P-lower and P-upper approximations of ($Cl_t^{\leq}$) are: $\underline{P}(Cl_t^{\leq}) = \{x \in \cup: D_P^-(x) \subseteq Cl_t^{\leq}\}$, $\overline{P}(Cl_t^{\leq}) = \bigcup_{x\in Cl_t^{\leq}} D_P^-(x)$, for *t*=1,...,n. The P-boundary sets of $Cl_t^{\geq}$ and $Cl_t^{\leq}$ are: $Bn_P(Cl_t^{\geq}) = \overline{P}(Cl_t^{\geq}) - \underline{P}(Cl_t^{\geq})$, $Bn_P(Cl_t^{\leq}) = \overline{P}(Cl_t^{\leq}) - \underline{P}(Cl_t^{\leq})$, for *t*=1,...,n. These approximations to the unions of "upward" and "downward" decision classes can be used to infer decision rules of the form "*if ... then ...*". For a given union of "upward" or "downward" of decision classes $Cl_t^{\geq}$ or $Cl_t^{\leq}$, *s,t* ∈ *T*, the rules induced under the hypothesis that objects pertaining to lower approximations $\underline{P}(Cl_t^{\geq})$ or $\underline{P}(Cl_t^{\leq})$ are positive and all others are negative suggest that an object be attributed to "at least one class Cl_t" or to "at most one class Cl_s", respectively. These rules are known as "certain decision rules" ($D_{\leq}$ or $D_{\geq}$) because they attribute objects to unions of decision classes without any ambiguity. Alternatively, if objects pertain to upper approximations, the rules are known as "possible decision rules"; thus, objects could pertain to "at least one class Cl_t" or "at most one class Cl_s". Finally, if objects pertain to the intersection $\overline{P}(Cl_s^{\leq}) \cap \overline{P}(Cl_t^{\geq})$ (*s*<*t*), the rules induced are known as "approximate rules", i.e., objects are between classes Cl_s and Cl_t. Therefore, if for each criterion $q \in C$, $V_q \subseteq \mathbf{R}$ (V_q is quantitative) and for each $x,y \in U$, $f(x,q) \geq f(y,q)$ implies xS_qy (V_q has a preferential ordering), decision rules can be considered to be of five types:

1- certain $D_{\geq}$-decision rules:
if $f(x,q_1) \geq r_{q1}$ *and* $f(x,q_2) \geq r_{q2}$ *and ...* $f(x,q_p) \geq r_{qp}$, *then* $x \in Cl_t^{\geq}$;
2- possible $D_{\geq}$-decision rules:
if $f(x,q_1) \geq r_{q1}$ *and* $f(x,q_2) \geq r_{q2}$ *and ...* $f(x,q_p) \geq r_{qp}$, *then x possibly belongs to* $Cl_t^{\geq}$;
3- certain $D_{\leq}$-decision rules:
if $f(x,q_1) \leq r_{q1}$ *and* $f(x,q_2) \leq r_{q2}$ *and ...* $f(x,q_p) \leq r_{qp}$, *then* $x \in Cl_t^{\leq}$;
4- possible $D_{\leq}$-decision rules:
if $f(x,q_1) \leq r_{q1}$ *and* $f(x,q_2) \leq r_{q2}$ *and ...* $f(x,q_p) \leq r_{qp}$, *then x possibly belongs to* $Cl_t^{\leq}$, where $P = \{q_1, ..., q_p\} \subseteq C$, $(r_{q1}, ..., r_{qp}) \in V_{q1} \times V_{q2} \times ... \times V_{qp}$ and $t \in T$;
5- approximate $D_{\leq\geq}$-rules:
if $f(x,q_1) \geq r_{q1}$ *and* $f(x,q_2) \geq r_{q2}$ *and ...* $f(x,q_k) \geq r_{qk}$ *and* $f(x,q_{k+1}) \leq r_{qk+1}$ *and* $f(x,q_p) \leq r_{qp}$, *then* $x \in Cl_s \cup Cl_{s+1} \cup ... \cup Cl_t$.

Rules of types "1" and "3" represent "certain knowledge" extracted from a data table (or information system), rules of types "2" and "4" represent "possible knowledge", and the rule of type "5" represent "ambiguous knowledge". As an example of the application of these preceding concepts, Table 2 contains a data table with three condition criteria C = {q_1, q_2, q_3}, all preferably maximised, and three decision classes Cl_1, Cl_2 and Cl_3, with preferential ordering in increasing numerical order [12].

Table 2. Data table with 3 condition criteria and 3 decision classes

Object	q1	q2	q3	d
1	1.5	3	12	Cl2
2	1.7	5	9.5	Cl2
3	0.5	2	2.5	Cl1
4	0.7	0.5	1.5	Cl1
5	3	4.3	9	Cl3
6	1	2	4.5	Cl2
7	1	1.2	8	Cl1
8	2.3	3.3	9	Cl3
9	1	3	5	Cl1
10	1.7	2.8	3.5	Cl2
11	2.5	4	11	Cl2
12	0.5	3	6	Cl2
13	1.2	1	7	Cl2
14	2	2.4	6	Cl1
15	1.9	4.3	14	Cl2
16	2.3	4	13	Cl3
17	2.7	5.5	15	Cl3

The unions of classes are as follows:
$Cl_1^{\leq}$ = {3,4,7,9,14}; $Cl_2^{\leq}$= {1,2,3,4,6,7,9,10,11,12,13,14,15}; $Cl_2^{\geq}$= {1,2,5,6,8,10,11,12,13,15,16,17}; $Cl_3^{\geq}$= {5,8,16,17}.

There are 5 objects that violate the Dominance principle: 6, 8, 9, 11 and 14. For example, object "9" dominates object "6" because it is better in all condition criteria (q_1, q_2 and q_3). However, it belongs to decision class Cl_1, worse than Cl_2. Next, upper and lower approximations of each decision class were computed:
$\underline{C}(Cl_1^{\leq})$ = {3, 4, 7}; $\overline{C}(Cl_1^{\leq})$ = {3, 4, 6, 7, 9, 14}; $\underline{C}(Cl_2^{\leq})$ = {1, 2, 3, 4, 6, 7, 9, 10, 12, 13, 14, 15}; $\overline{C}(Cl_2^{\leq})$ = {1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}; $\underline{C}(Cl_2^{\geq})$ = {1, 2, 5, 8, 10, 11, 12, 13, 15, 16, 17}; $\overline{C}(Cl_2^{\geq})$ = {1, 2, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17}; $\underline{C}(Cl_3^{\geq})$ = {5, 16, 17}; $\overline{C}(Cl_3^{\geq})$ = {5, 8,11, 16, 17}.

Following the analysis sequence proposed in the DOMLEM algorithm [12] regarding rules of type "1", we extracted the decision rules and the respective objects satisfying those rules and their evaluation metrics - ($[e_i] \cap G/[e_i]$) and ($[e_i] \cap G$), where "e_i" represents a rule and "G" represents the upper approximation under analysis – $\underline{C}(Cl_3^{\geq})$:

e_1= ($f(x,q_1) \geq 2.3$), {5, 8, 11,16, 17}, 0.6 , 3; e_2= ($f(x,q_1) \geq 2.7$), {5, 17}, 1.0 , 2;
e_3= ($f(x,q_2) \geq 4$), {2, 5, 11, 15, 16, 17}, 0.5 , 3; e_4= ($f(x,q_2) \geq 4.3$), {2, 5, 15, 17}, 0.5 , 2;
e_5= ($f(x,q_2) \geq 5.5$), {17}, 1.0 , 1; e_6= ($f(x,q_3) \geq 9$), {1, 2, 5, 8, 11, 15, 16, 17}, 0.38 , 3;
e_7= ($f(x,q_3) \geq 13$), {15, 16, 17}, 0.67 , 2; e_8= ($f(x,q_3) \geq 15$), {17}, 1.0 , 1.

Decision rule e_2 is chosen, given that it has the highest value for the evaluation metric (1.0) and more objects (2) in the "$[e_i] \cap G$" intersection, aside from satisfying condition "$[e_2] \subseteq B$". These objects are then excluded from G, and the same procedure to extract decision rules is applied to the remaining object ("16"). The rules then inferred are:

e_9 = ($f(x,q_1) \geq 2.3$), {8, 11, 16}, 0.33 , 1; e_{10}= ($f(x,q_2) \geq 4$), {2, 11, 15, 16}, 0.25 , 1;
e_{11}= ($f(x,q_3) \geq 13$), {15, 16}, 0.5 , 1.

Rule e_{11} has the highest evaluation metric value (0.5), but because object "15" does not belong to the approximation being analysed ($\underline{C}(Cl_3^{\geq})$), one must then infer "complex" rules ("^"): $e_9 \wedge e_{11}$ and $e_{10} \wedge e_{11}$. Therefore, rule $e_9 \wedge e_{11}$ is chosen because it has the highest evaluation metric value and covers the lower approximation's elements. Taking only the lower approximation to decision class Cl_3 into consideration, the following minimal set of decision rules is obtained:

if ($f(x,q_1) \geq 2.7$), *then* $x \in Cl_3^{\geq}$ {5, 17};
if ($f(x,q_1) \geq 2.3$) *and* ($f(x,q_3) \geq 13.0$), *then* $x \in Cl_3^{\geq}$ {16, 17}.

A generalisation for DRSA has been proposed, called VC-DRSA (*Variable consistency-DRSA*) [12], [13], which allows one to define lower approximations to unions of decision classes that take a limited number of negative examples controlled by a predefined "consistency level" $l \in (0, 1]$. In VC-DRSA, each decision rule is characterised by an additional parameter "α" known as the rule's "confidence" (level). Some of its basic concepts are as follows: a rule's “strength” is the ratio of the number of objects that satisfy the rule to the total number of objects, its “certainty” is the ratio of the number of objects that satisfy the rule to the number of objects that satisfy the rule's condition criteria, and its “coverage” is the ratio of the number of objects that satisfy the rule to the number of objects that satisfy the rule's decision criteria. The coverage factor is the estimate of conditional probability that Φ is true in S given Ψ is true in S, with the probability [7], [14]:

$$\mathrm{cov}_s(\Phi \mid \Psi) = \mathrm{card}(\|\Phi \wedge \Psi\|_s) \,/\, \mathrm{card}(\|\Psi\|_s) \quad (4)$$

4. Application of the Dominance principle to Web content search - specific case

This study originated from the Web search by "Rio de Janeiro" (City and/or State, in Brazil). But in return, there were over 340 million results (URLs) - based "16-feb-2016", by "Google". Thus, for the result of the research was the most effective and restricted, were added a few words to the search. Considered in this study, search criteria or "condition": beach, football, samba, show, restaurant, museum, exhibition, theater. In all, they were considered nine search criteria (including "rio de janeiro"); each was separated by the logical connective "and" to make clear a desirable outcome is one contained if possible. The search engine returned approximately 468,000 results, and itself was limited to return the URLs more relevant - in this case, 96 results. This text was then exported to a spreadsheet (Microsoft Excel) and previously treated by an algorithm in VBA (Visual Basic for Applications). This algorithm aimed to tabulate the citation frequency of each condition criterion which appeared in each summary of text (from URL). At this table with the condition criteria (except the URL, last column), was added to the "ranking" of URLs (returned by the search engine). To make it possible to build the decision table, it was included an "information class". The information class was established as follows: split the universe (or "ranking") of URLs in three parts (approximately) equal: the first part is the value of information class "1"; the intermediate part, the value of information class "2"; and the end part, the value of information class "3". Among the information classes, we establish the relation of "strict preference" ("**P**") information class "1" is better than the information class "2" ($Cl_1 \mathbf{P} Cl_2$) which, in turn, is better than the information class "3" ($Cl_2 \mathbf{P} Cl_3$). The “information class” was considered like "cost" - the lower the value, the better. In this study, the criterion "ranking" returned by the search engine was only used as a reference ("neutral"). The condition criteria were considered like "gain" - the higher the value, the better. The data tabulated were then subjected to an analysis by the Dominance principle (DRSA), to identify the URL whose sites contained search criteria (or condition) with maximum values, and the value of information class was minimal (Table 3, with the first 20 URLs), considering the repetition of condition criteria. In this case, the software jMAF showed a "core" of condition criteria: "rio de janeiro, beach, football, samba, show, exhibition, theater".

Table 3. Decision table with nine condition criteria and a decision criterion (the first 20 URLs, total 96 URLs)

ranking	rio de janeiro	beach	football	samba	show	restaurant	museum	exhibition	theater	information class	URL (Universal Resource Locator)
1	1	1	0	1	1	0	1	1	0	1	vejario.abril.com.br/materia/eventos/programacao-450-anos-rio
2	1	0	0	0	0	0	0	0	1	1	guiadeniteroi.com/
3	1	1	0	0	1	0	0	1	0	1	guia.uol.com.br/rio-de-janeiro/shows/.../rock-in-rio-veja-como-chegar-p...
4	1	0	0	1	1	0	0	0	0	1	guia.uol.com.br/rio-de-janeiro/shows/.../sesc-celebra-o-dia-do-comerciari...
5	1	2	0	0	1	0	0	0	0	1	guia.uol.com.br/rio-de-janeiro/shows/detalhes.htm?ponto=o...praia...
6	1	0	1	1	1	1	0	0	0	1	https://degracaeuvou.wordpress.com/
7	0	1	1	0	1	1	1	1	1	1	www.guiadasemana.com.br/turismo/noticia/programacao-gratis-em-sp
8	1	0	1	1	1	0	0	1	1	1	zonanorteetc.com.br/
9	2	1	1	1	1	1	1	0	1	1	comsut.com.br/wp/links/
10	1	2	1	0	0	1	1	1	1	1	www.acesseimovel.com.br/Rio_de_Janeiro_Pontos_Turisticos_Praias_Ho...
11	1	1	1	1	1	1	1	1	1	1	www1.uol.com.br/bibliot/turismo/riojancp.htm
12	2	1	0	1	1	0	0	0	1	1	www.blogsoestado.com/pedrosobrinho/
13	1	2	1	0	1	1	1	0	0	1	guia.melhoresdestinos.com.br/o-que-fazer-rio-de-janeiro-4-20-p.html
14	0	1	0	1	2	1	0	1	1	1	www.riolight.com.br/tag/rio-de-janeiro/page/48/
15	0	0	1	0	1	1	1	1	1	1	musikcity.mus.br/ra/mixfmcuritiba_main.html
16	1	1	1	0	0	0	0	1	1	1	https://pt.wikibooks.org/...Rio_de_Janeiro.../Primeira_metade_do_século...
17	1	1	0	0	1	2	1	0	1	1	windsorhoteis.com/conheca-o-rio/
18	1	1	0	0	0	1	1	0	1	1	https://catracalivre.com.br/rio/lugares/
19	1	0	0	1	1	0	0	1	1	1	postozero.com/eventos
20	1	0	0	1	0	1	1	0	1	1	www.elmistihouse.com/agenda-de-atividades-no-rio-de-janeiro

The decision rules were then generated by software jMAF, using the Dominance principle (DRSA) - Table 4:

Table 4. Rules (9) generated by software jMAF

<table>
<tr><th colspan="2">Rules</th></tr>
<tr><td>1: (rio_de_janeiro >= 1) & (show >= 1) & (theater >= 2) => (information_class <= 1) |CERTAIN, AT_LEAST, 1|
LearningPositiveExamples: 9, 10, 13, 14, 21, 25
Support: 2
SupportingExamples: 21, 25
Strength: 0.021052631578947368
Confidence: 1.0
CoverageFactor: 0.06451612903225806
Coverage: 2
CoveredExamples: 21, 25</td><td>2: (beach >= 2) & (football >= 1) & (show >= 1) => (information_class <= 1) |CERTAIN, AT_LEAST, 1|
LearningPositiveExamples: 9, 10, 13, 14, 21, 25
Support: 1
SupportingExamples: 13
Strength: 0.010526315789473684
Confidence: 1.0
CoverageFactor: 0.03225806451612903
Coverage: 1
CoveredExamples: 13</td></tr>
<tr><td>3: (beach >= 2) & (football >= 1) & (exhibition >= 1) => (information_class <= 1) |CERTAIN, AT_LEAST, 1|
LearningPositiveExamples: 9, 10, 13, 14, 21, 25
Support: 1
SupportingExamples: 10
Strength: 0.010526315789473684
Confidence: 1.0
CoverageFactor: 0.03225806451612903
Coverage: 1
CoveredExamples: 10</td><td>4: (show >= 2) & (exhibition >= 1) => (information_class <= 1) |CERTAIN, AT_LEAST, 1|
LearningPositiveExamples: 9, 10, 13, 14, 21, 25
Support: 1
SupportingExamples: 14
Strength: 0.010526315789473684
Confidence: 1.0
CoverageFactor: 0.03225806451612903
Coverage: 1
CoveredExamples: 14</td></tr>
<tr><td>5: (rio_de_janeiro >= 2) & (beach >= 1) & (football >= 1) & (show >= 1) => (information_class <= 1) |CERTAIN, AT_LEAST, 1|
LearningPositiveExamples: 9, 10, 13, 14, 21, 25
Support: 1
SupportingExamples: 9
Strength: 0.010526315789473684
Confidence: 1.0
CoverageFactor: 0.03225806451612903
Coverage: 1
CoveredExamples: 9</td><td>6: (samba >= 1) & (exhibition >= 2) => (information_class <= 2) |CERTAIN, AT_LEAST, 2|
LearningPositiveExamples: 9, 10, 12, 13, 14, 21, 25, 32, 37, 47, 48, 52
Support: 2
SupportingExamples: 47, 48
Strength: 0.021052631578947368
Confidence: 1.0
CoverageFactor: 0.03225806451612903
Coverage: 2
CoveredExamples: 47, 48</td></tr>
<tr><td>7: (rio_de_janeiro >= 2) & (beach >= 1) & (show >= 1) => (information_class <= 2) |CERTAIN, AT_LEAST, 2|
LearningPositiveExamples: 9, 10, 12, 13, 14, 21, 25, 32, 37, 47, 48, 52
Support: 4
SupportingExamples: 9, 12, 37, 52
Strength: 0.042105263157894736
Confidence: 1.0
CoverageFactor: 0.06451612903225806
Coverage: 4
CoveredExamples: 9, 12, 37, 52</td><td>8: (show >= 1) & (theater >= 2) => (information_class <= 2) |CERTAIN, AT_LEAST, 2|
LearningPositiveExamples: 9, 10, 12, 13, 14, 21, 25, 32, 37, 47, 48, 52
Support: 3
SupportingExamples: 21, 25, 32
Strength: 0.031578947368421054
Confidence: 1.0
CoverageFactor: 0.04838709677419355
Coverage: 3
CoveredExamples: 21, 25, 32</td></tr>
<tr><td>9: (beach >= 1) & (show >= 2) => (information_class <= 2) |CERTAIN, AT_LEAST, 2|
LearningPositiveExamples: 9, 10, 12, 13, 14, 21, 25, 32, 37, 47, 48, 52
Support: 2
SupportingExamples: 14, 52
Strength: 0.021052631578947368
Confidence: 1.0
CoverageFactor: 0.03225806451612903
Coverage: 2
CoveredExamples: 14, 52</td><td></td></tr>
</table>

Selecting the rules that aim (necessarily) the presence of the condition criterion "rio de janeiro", and other condition criteria to the highest possible value, and concomitantly with the decision criterion "information class" to the lowest possible value, there is obtained the following Table 5:

Table 5. Rules necessarily with the presence of the condition criterion “rio de janeiro”

# Rule	Rule	# URL (SupportingExamples)
1	(rio_de_janeiro >= 1) & (show >= 1) & (theater >= 2) => (information_class <= 1)	21, 25
5	(rio_de_janeiro >= 2) & (beach >= 1) & (football >= 1) & (show >= 1) => (information_class <= 1)	9
7	(rio_de_janeiro >= 2) & (beach >= 1) & (show >= 1) => (information_class <= 2)	9, 12, 37, 52

Selecting up manually in spreadsheet with the condition criterion "rio de janeiro" greater than or equal to "1", "show" greater than or equal to "1", "theater" greater than or equal to "2" and "information class" less than or equal to "1"- rule "1", it has (Table 6):

Table 6. URLs aimed by rule “1”

ranking	rio de janeiro	beach	football	samba	show	restaurant	museum	exhibition	theater	information class	URL (Universal Resource Locator)
21	1	1	0	0	1	0	1	1	2	1	www.riodejaneironow.com/cultura.htm
25	1	0	1	1	1	1	0	1	2	1	revistatrustme.com.br/rio-de-janeiro-destinocopa2014-copa2014-turismo...

Selecting up manually in spreadsheet with the condition criterion "rio de janeiro" greater than or equal to "2", "beach" greater than or equal to "1", "football" greater than or equal to "1", "show" greater than or equal to "1" and “information class" less than or equal to "1"- rule "5", it has (Table 7):

Table 7. URLs aimed by rule “5”

ranking	rio de janeiro	beach	football	samba	show	restaurant	museum	exhibition	theater	information class	URL (Universal Resource Locator)
7	0	1	1	0	1	1	1	1	1	1	www.guiadasemana.com.br/turismo/noticia/programacao-gratis-em-sp
9	2	1	1	1	1	1	1	0	1	1	comsut.com.br/wp/links/
11	1	1	1	1	1	1	1	1	1	1	www1.uol.com.br/bibliot/turismo/riojancp.htm
13	1	2	1	0	1	1	1	0	0	1	guia.melhoresdestinos.com.br/o-que-fazer-rio-de-janeiro-4-20-p.html

The previous Table 7 shows for the condition criterion "rio de janeiro" greater than or equal to "2", kept the other condition criteria of the rule "5", there is only one URL that attends to this rule: URL "9". Selecting up manually in spreadsheet with the condition criterion "rio de janeiro" greater than or equal to "2", "beach" greater than or equal to "1", "show" greater than or equal to "1" and "information class" less than or equal to "2"- rule "7", it has (Table 8):

Table 8. URLs aimed by rule “7”

ranking	rio de janeiro	beach	football	samba	show	restaurant	museum	exhibition	theater	information class	URL (Universal Resource Locator)
9	2	1	1	1	1	1	1	0	1	1	comsut.com.br/wp/links/
12	2	1	0	1	1	0	0	0	1	1	www.blogsoestado.com/pedrosobrinho/
37	2	1	0	1	1	1	1	0	1	2	www.viagemja.com/blog/destinos/nacionais/rio-de-janeiro-2/
52	2	2	0	1	2	1	0	0	0	2	www.gohouse.com.br/servicos/

By the previous Tables 6, 7 and 8, it follows that, rules "1", "5" and "7" allows selecting those URLs with the highest possible values for the condition criteria, especially condition criterion "rio de janeiro", considering the decision criterion "information class" to the lowest possible value.

From the Coverage factor about the rules “1” and “7”, we get the following characterization about the URLs (“inverse algorithm”):

• 6.45 % with “information class” less than or equal to "1" have “rio de janeiro” greater than or equal to "1" and “show” greater than or equal to "1" and “theater” greater than or equal to "2" - rule “1”;

• 6.45 % with “information class” less than or equal to "2" have “rio de janeiro” greater than or equal to "2" and “beach” greater than or equal to "1" and “show” greater than or equal to "1" - rule “7”.

Thus, the previous Table 5 showed the possibility to extract "rules" about the set of URLs, using a "core" of condition criteria ("rio de janeiro, beach, football, samba, show, exhibition, theater") from the decision table.

5. Conclusions and recommendations for future work

In the context of this study, the search for "Rio de Janeiro" followed by eight other words, considered "condition criteria", exemplified a case of web content search. Adding condition criteria made it possible to obtain a more effective result and restricted. But still, the amount of URLs returned is significant (approximately 468,000). How to make the search results more effective? From the unstructured data that were returned by the search engine, it has become feasible to draw up a table with structured data, through the lifting of the citation frequency of condition criteria for each referenced URL summary. At this table, it was associated with a decision class ("information class"), where it was possible to expand it to a "decision table". Subsequently, the decision table associated to Dominance principle, which allow extracting "patterns" (or rules) and hence add information to "ranking" of URLs. In this case, a "core" of suggested condition criteria emphasized the importance in highlighting that subset of criteria that are essential to the information system (decision table) in the study, which could not be eliminated without impact (negative) to the system [8]. Of the 96 relevant URLs suggested by the search engine ("Google"), it is observed that the best positioned URLs do not always return the desired information – ex, the site referring to the URL "21" (www.riodejaneironow.com/cultura.htm) suggested by Rule "1", shows as much as or more information about "Rio de Janeiro" than the site referring to the URL "1" (vejario.abril.com.br/materia/eventos/programacao-450-anos-rio). About the significant URLs in the form of "ranking", the search engine according to its own criteria, exemplified in these cases, as it may become costly to attempt to analyze manually, a considerable mass of unstructured text. Thus, the logical rules generated based on a "decision table", allowed reveal patterns on the set of URLs returned by the search engine, however the existence of other tools and decision support techniques on "web mining" and in particular under uncertainties – ex, "document clustering" and "web mining soft" [15], [16]; "rough association rules" [17]; "rough-fuzzy" and "rough-wavelet" [18]. Futhermore, in statistical data analysis based on Bayes' Theorem, we assume that prior probability about some parameters without knowledge about the data is given. The posterior probability is computed next, which tells us what can be said about prior probability in view of the data. In the Rough Set approach the meaning of Bayes' Theorem is unlike. It reveals some relationships in the database, without referring to prior and posterior probabilities, and it can be used to reason about data in terms of approximate (rough) implications. It identifies probabilistic relationship between conditions and decisions in decision algorithms and can be used to give explanation (reasons) for decisions [10], [11]. By the way, the attempt in unifying logic and probability to logical sentences is shown in [19]. And for data mining applications for example, the acquisition of probabilistic, rather than deterministic, predictive models is of primary importance [20]. As a proposal for future study, the application of the Dominance principle in the generation of a "ranking" complementary to the original ranking, using the jRank software (Ranking using Dominance-based Rough Set Approach) [21].

Acknowledgements

The work leading up to this paper was partially supported by CNPq/Brazil, through Research Project n. 305732/2012-9. The authors thank the Institute of Computer Science of the Poznan University of Technology, Poland, for allowing them to use the jMAF software in their research.

References

[1] Pinheiro, C. A. R. Inteligência analítica – Mineração de dados e descoberta do conhecimento. Rio de Janeiro: Ed. Ciência Moderna;

2008.
[2] Dhar, V. Data science and prediction. Communications of the ACM, 2013, v. 56, n. 12, 64-73.
[3] Brin, S., Page, L. The anatomy of a large-scale hypertextual web search engine. Proceedings of the Seventh International Conference on World Wide Web 7, 1998, 107-117.
[4] Su, A., Hu, Y. C., Kuzmanovic, A., Koh, C. How to improve your google ranking: myths and reality. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, 50-57.
[5] Zadeh, L.A. Fuzzy sets. Information and Control, 1965, 8, p. 338–353.
[6] Blaszczynski, J., Greco, S., Matarazzo, B., Slowinski, R., Szeląg, M. jMAF. Institute of Computer Science. Poznan University of Technology, Poland. Available from: http://www.cs.put.poznan.pl/jblaszczynski/Site/jRS.html, 2012.
[7] Pawlak, Z. Rough sets and decision analysis. Information Systems & Operational Research, 2000, 38, 3, p. 132-144.
[8] Pawlak, Z. Rough sets. Theoretical aspects of reasoning about Data. Kluwer Academic Publishers, Dordrecht, 1991.
[9] Pawlak, Z., Slowinski, R. Rough set approach to multiattribute decision analysis. European Journal of Operational Research. Invited Review, 1994, 72, p. 443-459.
[10] Slowinski, R., Greco, S., Matarazzo, B. Rough set and rule based multicriteria decision aiding. Pesquisa Operacional, 2012, 32, 2, p. 213-269.
[11] Kotlowski, W., Slowinski, R. On nonparametric ordinal classification with monotonicity constraints, IEEE Transactions on Knowledge and Data Engineering, 2013, 25, 11, p. 2576-2589.
[12] Greco, S., Matarazzo, B., Slowinski, R., Stefanowski, J. An algorithm for induction of decision rules consistent with dominance Principle. In: Ziarko, W., Yao, Y. Rough Sets and Current Trends in Computing, LNAI. Springer-Verlag, Berlin, 2001, 2005, p. 304–313.
[13] Blaszczynski, J., Slowinski, R., Szelag, M. VC-DomLEM: Rule induction algorithm for variable consistency rough set Approaches. Research Report RA-07/09. Poznań University of Technology, 2009.
[14] Pawlak, Z. Rough sets, decision algorithms and bayes' theorem. European Journal of Operational Research, 2002, 136, 181-189.
[15] Pal, S. K., Talwar, V., Mitra, P. Web mining in soft computing framework: relevance, state of the art and future directions, IEEE Transactions on Neural Networks, 2002, v. 13, n. 5, 1163-1177.
[16] Nayak, R. K., Mishra, D., Das, S., Shaw, K., Mishra, S., Tripathy, R. Clustering and classifying informative attributes using rough set Theory. ACM International Conference on Advances in Computing, Communications and Informatics, 2012, 118-123
[17] Li, Y., Zhong, N. Rough association rule mining in text documents for acquiring web user information needs. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2006.
[18] Meher, S. K., Pal, S. K., Dutta, S. Granular computing models in the classification of web content data. IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2012, 175-179.
[19] Russell, S. Recent developments in unifying logic and probability. Communications of the ACM, July 2015, 58, 7, 88-97.
[20] Slezak, D., Ziarko, W. The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning, 2005, 40, 81-91.
[21] Szelag, M., Slowinski, R., Greco, S., Blaszczynski, J., Wilk, S. jRank. Institute of Computer Science. Poznan University of Technology, Poland: Available from: http://: < http://www.cs.put.poznan.pl/mszelag/Software/jRank/jRank.html>, 2013.