International
Journal of
**Human
Sciences
Research**

RECOVERY AND FILTERING OF HEALTH AND ENVIRONMENTAL TEXTS THROUGH INTERNET SEARCH MECHANISMS

Arno Rieder
Mato Grosso State University -UNEMAT
Cáceres, Mato Grosso, Brasil
https://orcid.org/0000-0002-9027-2549

Fabiana Aparecida Caldart Rodrigues
Mato Grosso State University -UNEMAT
Cáceres, Mato Grosso, Brasil
https://orcid.org/0000-0002-4293-3402

Abstract: This is a work of refining and testing methodology for search and retrieval of information from scientific documents on the Internet. The work involves Health and Environment, aiming to recover scientific studies available and accessible on the topic medicinal plants used for diabetes. We chose to use the search engine developed by Google Scholar (GA), entering in your search box key words (scientific name, diabetes, abstract) directing the search anywhere in the document (M_6) or in the ad text *"all in the title"* (M_5), or *"in the title"* ($M_{1,2,3,4}$). This resulted in the application of 6 search modes ($M_{1,2,3,4,5,6}$). Searches were made in December 2012 to 135 species of plants. Results reveal that recovery capabilities and filter models applied are significantly different. The less specific the search procedure applied more documents are retrieved, although the utilization is less the absence of linkage between the foci of interest (plants for diabetes). The opposite of this trend occurs on occasions as more specific instructions are the search procedure adopted, ie, there will be less recovered texts but already highly filtered, taking advantage of almost all.
Keywords: Online sources, Medicinal plants, Diabetes.

INTRODUCTION

More and more information, including scientific studies, is made available on the World Wide Web (WWW or Web) and, interconnected and executed on the World Wide Web (Internet). This virtual universe continues to improve and expand.

The volume of data, vehicles and scientific works made available on the Web is exponentially growing, in addition to the expansion of the scope and diversity of content (BIOLCHINI and GIORDANO, 2011). Paganelli (2011) attributes to content providers the responsibility for what they make available on the internet. Industry regulators and those who give access to the Internet to third parties must assume their share of responsibility for the consequences. The selection of content and the interpretation of available data must be the responsibility of the actors involved and their tutors. Nogueira (2012) in his study assigns the school responsibility for the procedures of its students in this process.

It is necessary to invest more in the training of online teachers in this era of cyber culture, as the manipulation of the digital world demands expert and politically sensible instructors (ROSSINI and SANTOS, 2013). Locuses are developed to store and offer different categories of data, information and knowledge, such as digital repositories(RDs).

RDs are *online* databases that group and organize the scientific production of an institution or thematic area (IBICT, 2014). Coelho (2014) claims that the RDs are formed by digitized documents or electronic files, on the Internet, allowing remote access. Facilitates selection, information distribution, preservation of document integrity. Remote access is via a networked computer; with the possibility of simultaneous users, including integrating several audio-visual resources (text, sound, image, etc.). They provide access to information and content capture in various ways from querying, copying, downloading, distributing, printing, searching or referencing. Technologies and procedures in the construction and management of digital collections are essentially dealt with by Information Science and Library Science (LEMOS, 2009).

In the area of health, there is the Virtual Health Library-VHL (CAMARA et al., 2008) which captures and gathers the dynamic production in this area. Lamas (2003/2004) alludes that virtual libraries, via the Internet, allow for the dissemination of more knowledge. In general, they provide access

to journals and/or make available selected and well-qualified scientific productions. The CAPES Journal Portal is considered good for searching technical and scientific documents.

The CAPES Journal Portal is an excellent vehicle for searching technical-scientific texts present in journals in the databases of various areas of knowledge, providing access to academia and researchers (LAMAS, 2003/2004). Accessing the abstract alone already allows you to assess the relevance of the document and the convenience of consulting the full text. The use of several databases allows for more scope and interdisciplinarity. A database is a set of structured and manipulated information through a Database Management System (FERNANDES et al., 2013).

Literature review is a methodological foundation for scientific investigation. The internet records human knowledge and provides access to content, sometimes limiting it by certain conditions to be met.

Camara *et al.* (2008) cites that scientific journals are important sources of knowledge dissemination. The texts accessible *online*, available on the internet, can be searched using appropriate descriptors and resources to retrieve them.

According to Sales and Vieira (2007), information retrieval from the virtual environment is affected by the evolution of information and communication technologies. The volume of production and dissemination of knowledge grows and the speed of circulation needs to adapt, as it can affect efficient recovery. They observed that research in information retrieval in Brazil was on the rise in 2007, although very concentrated in Computer Science and Information Science groups, but arousing the interest of scholars from other areas. The trend is to generate intelligent systems that enable interaction between people and information, with structured tools for document recovery.

Specific models of information retrieval systems or adjustments to them may be necessary for a certain area of knowledge, adapting to their peculiarities, focuses or interests, in order to retrieve documents at an opportune time. Precision and effectiveness are desired, as is the case with a proposition by Medeiros (2008), who emphasizes that it is up to the librarian, as a manager, to offer useful, accurate and timely instruction to their information demanders. Biolchini and Giordano (2011) claim that informal learning is still the main way to acquire skills in information search and retrieval, adopting more generic search engines and less advanced resources, with low use of operators and terms. It is necessary to train more to know how to operate the search and recovery.

Google Scholar (GA) has good database coverage (MARTINS, 2011). The GA gives visibility to publications, as any academic document published on the web becomes accessible (MUGNAINI and STREHL, 2008). Pinto and Pontes Junior (2008) states that the GA is viable for analyzing scientific information, comparable to that used by the ISI (Institute for Scientific Information) databases and; retrieval and indexing strategies are more reliable in GA; it also offers the citation index, expanding the scope of the researcher's interests. Caregnato (2011) recalls that GA came into existence on the Web at the end of 2004 to be a search tool for academic information and, currently, it retrieves a large number of documents surpassing traditional citation rates. He suggests that the mechanism needs improvement to improve the quality of the selection of data sources and records. Several investigations in online sources choose to use the Google search engine (Fernandes et al., 2005; Pinheiro et.al., 2008). Recovery through GA or another search tool requires a minimum of knowledge from its users about it and the professionals who develop it

should provide better instructions on its use (SOUZA, 2009).

Although there are many electronic databases and tools for accessing and retrieving data, an analogy can still be drawn to looking for a needle in a huge, randomly growing haystack. The volume of content helps hinder the specificity of the retrieved data (BIOLCHINI et al., 2011). So the improvement of search tools for data recovery is necessary to find these needles. A specific theme requires choices and adjustments in the most general search procedures. It has been imposed that the search for specialized information online needs to be carried out in increasingly sparse space and time (*BIOLCHINI et al., 2011*). In order to achieve the desired targets, these procedures sometimes still need to be established. In the present work, the aim is to adjust a methodology that retrieves the relevant academic texts from the virtual space in the areas of health and environment and that address the focus: plants used for diabetes control.

GOALS

Retrieve studies, via the internet, dealing with plant species (nc) for diabetes control, available and accessible online.

Apply different formats and modalities of searching for data of interest on the internet, to:

List text quantities by species; Highlight the species that retrieved more and less texts; Identify the amount of documents retrieved by search modality;

Know the recovery and filtering power of the applied procedures.

HYPOTHESES

-There are studies dealing with plant species (nc) for the control of diabetes: for some plants, yes (they may have few, moderate, many documents), but for others, no.

It is possible to find them, according to the desired specificity, using appropriate formats, keywords and means.

METHODOLOGY
PROCEDURES FOR ONLINE RETRIEVAL OF THEMATIC DOCUMENTS

Google Scholar (GA) search tool was applied to retrieve documents.

According to GOOGLE (2014), GA is a tool to search academic literature in various disciplines and sources, whose documents searched are listed in order of academic relevance. The most relevant references are displayed at the beginning of the results page. Relevance takes into account the full text of the article, the author, the vehicle in which the article was published and the frequency with which it was cited in other academic publications.

The keywords "*scientific name*"("nc")" (formed by genus and species), "*diabetes*"("d"), "*abstract*"("a") were used. A document is represented by a set of terms, which can be keywords, and then the search system, based on the user's query, retrieves the documents that contain such representation. (CORRÊA, 2005).

The three words used each form a set-universe to be explored. There are, therefore, three sets to focus on. Those of interest to the objectives of this work are subsets of intersection between them, capable of being encompassed by the search resources of the GA - which form the fourth set involved. To improve the recovery of documents as well as the use among those retrieved, it is essential to improve search strategies, including during the exercise of the process itself. Bonfim (2009) mentions that the adjustments in retrieval strategies are defined as a function of the probability of relevance and non-relevance of the documents resulting from

the searches. In the search box [cx] of GA, these three keywords were inserted between "quotes", thus establishing the limit and the exact characters of the keywords. Without the "quotes", the search encompasses a broad exploratory environment, far beyond the intended targets.

To instruct the search in GA, in any case, the first keyword entered in the search box was "nc":"scientific name" and three formats were used to insert these Fi in[cx]:

F1) Not conditioned for the keywords to be in the title (they could be anywhere in the document) ie: not necessarily in the title; F2) Conditioned to be "*all in the title*"; F3) Conditioned for the first keyword to be "*in the title*";

The completions of the [box], in the three formats (De) generated six unique search modalities (M_i), and one mode was used for both F1 and F2 and four modes for F3, filled in, one at a time, as follows:

M_1) F3.1 ["*in the title*":"*scientific name*"]; M_2) F3.2 ["*in the title*":"*scientific name*" "*abstract*"]; M_3) F3.3 ["*in the title*":"*scientific name*" "*diabetes*"]; M_4) F3.4 ["*in the title*":"*scientific name*" "*diabetes*" "*abstract*"]; M_5) F2 ["*all in the title*":"*scientific name*" "*diabetes*"]; M_6) F1 ["*scientific name*" "*diabetes*" "*abstract*"].

The F1 format is an instruction to search for keywords anywhere in the academic document that refer to the species for diabetes.

The F1 format uses three keywords and constitutes the search mode M_6. It is revealed texts that have the keywords "*nc*", "*d*", "*a*" anywhere in the document. There is direct insertion, in the search box [ ], of the 3 keywords: ["*nc*"+"*d*"+"*a*"]. With the insertions of "*nc*" there is a recovery of documents dealing with the species; with "a", academic texts and; with "d", related to diabetes and, whose keywords can be anywhere in the document. So this search mode (M_6) captures academic documents, not necessarily focused on title content, that discuss species of interest, probably related to diabetes, that is: academic texts that cite the species and diabetes, with or without link.

The F2 format shows texts that have all keywords in the title. There is insertion, in the box, of ["*all in the title*":"*nc*" "*d*"], this being the search mode M_5. Because the instruction of the keywords must all be in the title, this procedure retrieves texts that focus on the species relating it to diabetes. Therefore, F2=M_5 specifically seeks documents focused on species and diabetes (both in the title), ie: species-specific texts for diabetes. It is the most selective procedure in relation to the others, but it can eventually recover documents without academic format.

The F3 format unveils specific jobs with the "*nc*" of the plant species in the title, having in any part of the document the other keywords: "a" and/or "d". The four combinations, in F3, of "*nc*", "*a*", "*d*" are:$M_1=$F3.1 ["*in the title*":"*nc*"]; $M_2=$ F3.2 ["*in the title*":"*nc*" "*a*"];$M_3=$F3.3 ["*in the title*":"*nc*" "*d*"];$M_4=$F3.4 ["*in the title*":"*nc*" "*d*" "*a*"].

In F3.1 and F3.2 they look for documents, academics (M_2:"*a*") or not (M_1), species-focused ("*nc*" in the title), but with a different approach: they are texts about the species (non-specific).

The next two F3s bring documents, academics (M_4: "*a*") or not (M_3), species-focused ("*nc*" in the title) and for diabetes (M_3 and M_4): are texts of the species for diabetes.

In F3 the species must be in the title and the other keywords anywhere in the document. With this, texts are recovered, in M_1, that focus on the species in the title, with no other associated keywords; in M_2, focus on the species in the title and are academic works; in M_3, focus on the species in the title and refer to diabetes anywhere in the document and; in M_4 focus on the species in the title, being academic papers and refer to diabetes

anywhere in the document. Therefore, in short, these modalities retrieve academic documents (M_2, M_4) or not necessarily (M_1, M_3) dealing with the species in the title, referring to (M_3 and M_4) or not necessarily (M_1 e M_2) to diabetes, anywhere in the document and, being (M_2 and M_4) or not (M_1 and M_3) academic papers.

Combining the possibilities of F1, F2 and F3 there are six search modes and there are four that use the keywords: *"nc"* and *"d"* {*"nc"* in the title (M_5, M_3 and M_4) or not (M_6) and *"d"* inserted (M_5) or no(M_6, M_3 e M_4) in the title}. These are the ones of essential interest for the present work, as they recover *"nc"* and *"d"*, present in the title or anywhere in the document, with or without the keyword: *"a"* - identifying academic format or not.

Each search mode displays a certain number of texts on the computer screen. These are recovered documents. They were available somewhere, accessible at the time, given the search instructions applied and electronic resources adopted. Its content in part or in whole, on screen, can be read freely or with restrictions. If the restrictions are not overcome the essence and parts of the content are hidden.

QUANTITIES OF RETRIEVED TEXTS (A_T) VERSUS SEARCH MODALITIES (M_I)

Once the results were in possession and counted, the number of retrieved texts were successively grouped (A_T) at 5; 4; 3 and 2 classes. To define them, text quantity classes were used, as follows: A=0; B=1; C= 2 ├ 11; D= 11├ 51; E= ≥ 51.

From these, the following combinations were considered to constitute the groups of:

A_{T5} - five classes (one combination): [(A,B,C,D,E)]; A_{T4} - four classes (four combinations): [(AB,C,D,E); (A,BC,D,E); (A,B,CD,E); (A,B,C,DE)]; A_{T3} - three

classes (six combinations): [(A,B,CDE); (A,BCD,E); (ABC,D,E); (A,BC,DE); (AB,C,DE); (AB,CD,E)]; A_{T2} - two classes (four combinations): (A,BCDE); (AB,CDE); (ABC,DE); (ABCD,E).

This totaled 15 text number class groupings (A_T) retrieved in searches.

On the other hand, six unique modalities were defined before (M_I) of text search.

These were made up of comparative combinations, in groups of:

-a)6 singular modalities (a grouping): ($M_{1,2,3,4,5,6}$);

b) 2 modalities (fifteen groupings in pairs): ($M_{1,2}$); ($M_{1,3}$); ($M_{1,4}$); ($M_{1,5}$); ($M_{1,6}$); ($M_{2,3}$); ($M_{2,4}$); ($M_{2,5}$); ($M_{2,6}$); ($M_{3,4}$); ($M_{3,5}$); ($M_{3,6}$); ($M_{4,5}$); ($M_{4,6}$); ($M_{5,6}$)

Thus, 16 groups were obtained that compare the modalities (A_M) search engine. The fifteen text quantity class groupings (A_T) were contrasted with the sixteen groupings of search modalities (A_M).

RELATIVE RECOVERY AND FILTERING POWERS

The relative power of text retrieval (PR), based on the resources and instructions applied, is expressed, as a percentage, by the following equation:

PR%=100(n^o_{Mi}/n^o_{Mr}) or PR= 100 x (n^o_{Mi}= Number of texts retrieved for the modality in question)/(n^o_{Mr}= Number of texts retrieved by the majority reference modality).

The relative power of filtering or discarding texts (PF) is expressed as a percentage by the following equation: PF%=(100-PR). This shows that PR% + PF%=100%, and it is a comparison of the modality in question with another reference.

The use of texts (relevant documents) was carried out through selection by careful reading of all retrieved documents. According to Ramos and Munhoz(2011) the relevance of

the texts depends on the interests of the user represented in the search box [ ].

RESULTS AND DISCUSSION
INITIAL CONSIDERATIONS

Good quality and relevant information research needs its creators and operators to be, in addition to outstanding scholars on the subject that they embrace, skilled in the processes of identifying informational demand, the selection and proper use of information sources and search tools, and the formulation analysis, validation, selection and interpretation of results, as emphasized by Biolchini et al. (2011).

TEXTS RETRIEVED IN THE APPLIED SEARCH MODALITIES

The retrieved texts or documents (DR=re+nr) are considered relevant (re) or not relevant (nr). The re are academic texts that address studies of the species in question, relating it to diabetes or any aspect related to the syndrome that involves it. They can be of direct link - when the objective of the work presented in the document relates the species to diabetes, or indirect link - when this relationship is addressed not directly by the authors, but through other referenced studies (bibliographic citation). Therefore, there are documents of two relevant circumstances (re=d+i)): direct and indirect. Relevant retrieved texts are used documents.

Non-relevant or unused texts are those that do not fit into any of the above circumstances, that is: their studies do not link the species to diabetes or any aspect related to the metabolic syndrome.

THE TEN SPECIES WITH THE HIGHEST NUMBER OF TEXTS DISPLAYED (TABLE 1)

These species are not necessarily the same in the six unique search modalities. The set of modalities formed a final list of 18 species by including, in each modality, the 10 with the most texts. However, the species that retrieve more documents do not necessarily provide greater use, that is, they may not gather a greater number of relevant texts. Variations in the rate of use of retrieved texts are also found in other studies (MACHADO DE MORAES and DIAS MIRANDA, 2011). Changes in the amount of documents retrieved can occur when using small variants in search terms, as also verified by Milanez (2011). The two modes that retrieve documents about the species *"in the title"* (M_1 and M_2) presented the same list of plants with the ten largest amounts of texts, including: M_1 and M_2, an inversion of position of the 4th and 6th and, 8th and 9th places in the order of classification of species in the lists. The same species among the top ten lists also occurred in the modalities that retrieve documents about the species "*in the title*" and that mention diabetes (M_3 and M_4), but with inversion of the 3rd and 5th places.

The modality pairs: M_1 and M_2 had in common with the pairs: M_3 and M_4 only four species from the list of the 10 predominant in texts, but these still in different placements.

The modality: M_5 comparing to M_3 and M_4 presented eight species in common among the list of the 10 with the highest number of texts, and only one in the same order of classification. These modalities (M_3, M_4 e $M_{5,}$) evidence, in the groupings, convergence towards the species that predominate in the amount of texts, although not necessarily in the same order.

From the list of the 10 with the most retrieved texts, M_6 includes six species and, which are part of $M_{5,}$ but seven are present in the modalities M_1, M_2, M_3 and M_4. Among the 10 with the most retrieved texts, there are three species in common (garlic, onion and neem) in the six search modalities.

Two species do not even appear among the 10 with the highest number of texts on the lists of M_1 and of M_2 (that do not include "d"), that is, they are not among the 10 species with the highest number of documents retrieved by the modality you are looking for. "nc". They are probably not among the ten most studied species of the 135 used to control diabetes in Mato Grosso (MT). The reverse is also true: the two plans with the most documents retrieved by the keyword "nc" [M_1 e M_2: milho (*Zea mays*) and rice (*Oryza sativa*)] nor do they appear in the top ten with the highest number of texts in the lists. of M_5 and M_3 e M_4, those dealing with species studies for diabetes.

SPECIES USED AS ANTIDIABETICS IN TM: HIGHLIGHTS IN RETRIEVED DOCUMENTS

Of the 10 most cited in the 91 works that provided the list of plants used to control diabetes in TM (Rieder, 2009), only three species are part of the 18 that make up the group of 10 with the most documents retrieved in the six unique search modalities (M_1;..;M_6). Although widely used for diabetes in TM, a minority of species are among those with more documents retrieved (Table 1). The amount of documents retrieved can change with the variation of descriptors, search procedures used and the targeted compartments, which is also verified in Rosa's study (2009).

SPECIES WITH THE HIGHEST NUMBER OF RETRIEVED TEXTS

In the search modalities that use the keywords "*scientific name*" and "*diabetes*", the melon: Melão-de-São-Caetano (*Mormodica charantia*) is the species with the most texts retrieved by M_3, M_4 e M_5 and aloe (*Aloe vera*) by M_6 and, which occupies the 2nd place in M_3, M_4 and M_5. The modalities that retrieve documents for the species, but do not include "*diabetes*" as keyword (M_1 e M_2), present the

corn (*Zea mays*) first of all.

TEXT ABSENCES FOR SPECIES ON DIABETES BY SEARCH MODALITY

The modality with the most generic instruction (M_6) presents fewer species with absence of texts about the plant for diabetes.

The slightly less specific modalities, M_4 and M_3 presented absence of texts, respectively for 51 and 52 species. But M_5 was the one that did not retrieve texts on the subject for the greatest number of species (104).

The absence of texts for species indicates plants not necessarily without studies on the subject. It can represent texts not available to the means of access due to the search instructions used or incapable of being retrieved by the adopted mechanism. However, over time relevant documents will likely exist and be retrieved.

Delicato *et al.* (2001) propose intelligent systems for the personalized filtering of information on the Web, whose agents learn through user feedback and refine their searches, obtaining better results over time.

In the six search modalities, the number of species without documents retrieved varied [M_1=12(8,88%); M_2=18(13,33%); M_3=51(37,77%); M_4=52(38,51%); M_5=104 (77,03%); M_6=5 (3,70%)]. M_5 presented more species without texts (104 out of 135) as the procedure required the keywords in the title (**"nc", "d"**).

MODALITIES WITH A MINOR ABSENCE OF DOCUMENTS ABOUT THE SPECIES

The modalities with the lowest number of species with absence of texts were M_6 ("nc", "d", "a" in the document), M_1 ("nc" in the title) e M_2("nc" in the title and "a" in the document). They are the least filtering. They covered, respectively, five (3.70%), 12 (8.88%) and 18 (13.33%) species. The less focused procedures

(M_6) and those who do not specify the word diabetes (M_1; M_2) they have the fewest species in the absence of texts. However, the recovered texts have a more different content than the one to be covered (plants vs diabetes), which generates a large amount of discards in the next filtering.

As Seibel Junior (2007) mentions, the ideal would be to have a system capable of retrieving all relevant documents without including irrelevant ones. You may want to maximize the recovery of relevant documents and minimize irrelevant returns. The differences between the procedures are evident in the number of relevant and irrelevant documents

TEXTS DISPLAYED IN SEARCH MODES

When comparing (Test: $x^2_{\alpha(GL)}$) text quantity groupings [A_T:(A,B,C,D,E); (AB,C,D,E); (A,BC,D,E); (A,B,CD,E); (A,B,C,DE); (A,B,CDE); (A,BCD,E); (ABC,D,E); (A,BC,DE); (AB,C,DE); (AB,CD,E); (A,BCDE); (AB,CDE); (ABC,DE); (ABCD,E)] within the search modality groupings [A_M: ($M_{1,2,3,4,5,6}$); ($M_{1,2}$); ($M_{1,3}$); ($M_{1,4}$); ($M_{1,5}$); ($M_{1,6}$); ($M_{2,3}$); ($M_{2,4}$); ($M_{2,5}$); ($M_{2,6}$); ($M_{3,4}$); ($M_{3,5}$); ($M_{3,6}$); ($M_{4,5}$); ($M_{4,6}$); ($M_{5,6}$)], it turns out that 15 A_T, the species proportions are similar (p>0,05) only in three pairs of search modes [$A_{M:}$ ($M_{1,2}$); ($M_{1,6}$); ($M_{3,4}$)]. In the grouping: $M_{2,6}$ there is similarity in the number of species in three categories of number of texts: A_T [(ABC,D,E); (ABCD,E) (ABC,DE)]. And also in $M_{2,6}$, nine A_T [(A,B,C,D,E); (AB,C,D,E); (A,BC,D,E); (A,B,CDE); (A,B,CD,E); (A,B,C,DE); (A,BC,DE); (AB,C,DE); (AB,CD,E)] showed significant differences (p<0,05), while three: A_T [(A,BCD,E); (A,BCDE); (AB,CDE)] highly significant differences (p<0.01) regarding the number of species in the crosses: A_T x M_i.

In the remaining 12 combinations, the differences in the number of species were highly significant (p < 0.01) in all 16 groupings of number of texts. Therefore, situations of dependence on the amount of texts retrieved for the species as a function of the search modalities predominated. Some topics searched provide more and better articles than others, which was also observed by Lamas (2003/2004).

STATISTICAL MEASURES (TABLE 3)

In all search modalities, the amount of texts captured has a much higher standard deviation (2 to 4 times) than the average, resulting in a coefficient of variation between 201.96% and 411.50%. This indicates a very high dispersion of data (number of retrieved texts) in any of the search modes.

This dispersion, in the most filtering modalities, is justified by the fact that there is a greater concentration of species with an absence or few texts and, on the other hand, few species with a very high amount of retrieved texts. This is demonstrated by the polygon type obtained for the frequency distribution of the text quantity classes. In the six search modes, the shape of the figure is leptocurtic (sharp, that is, with a high concentration of species in the left-hand region of few texts) and positive asymmetric (high concentration of species on the left and stretching to the right, where there are many texts, but for fewer and fewer species).

SEARCH MODALITIES WITH THE LOWEST AND HIGHEST % OF SPECIES IN THE CLASSES OF A_T

The smallest(-) and the largest(+) % of species in the document quantity classes (A_T) recovered (A,B,C,D,E,) are registered in the search modes, as follows: A(No text):[-M_6(3,70%); +M_5(77,03%)]; B(a single text):[-M_1(2,96%); +M_4(8,88%)]; C(Two to ten texts):[-M_1(11,85%); +M_3(22,22%)]; D(Eleven to fifty texts):[-M_5(2,22%); +M_2(27,40%)]; E(More than fifty texts):[-M_5(0,74%); +M_1 and M_6 (53,33%)].

| Plants used to control diabetes(db) in MT (Rieder, 2009) (out of 135) with the highest number of texts displayed (ten more) by search procedures adopted | | | Bibliographies (of 91 that cite them p/db in MT (Number of sources)) | Search target in document (F1:anywhere =M6, F2: "all in the title"=M5; F3: "in the title"=M1,2,3,4,) and combination of keywords ["scientific name" (nc), "diabetes(d), "abstract"(a)] instructed in the GA search box and, respective, numbers of retrieved texts (displayed on the screen) for the targeted species | | | | | | sum in the lines |
| | | | | F3("in the title":"nc") | | | | | | |
Common name	Botanical family	Scientific name		M1="nc"	M2="nc"+"a"	M3="nc"+"d"	M4="nc"+"a"+d	F2=M5 ("all in the title": "nc" "d")	F1=M6 ("nc" "d" "a")	Texts
Cipo tripa-	Fabaceae	Bauhinia longipetala	1	0	0	0	0	0	0	0
Pata-de-	Fabaceae	Bauhinia rubiginosa Bong.	1	0	0	0	0	0	0	0
Insulina	Vitaceae	Cissus alata Jacq.	2	0	0	0	0	0	0	0
Carobinha	Bignoniace	Jacaranda semiserrata	1	0	0	0	0	0	0	0
Erva-de-	Loranthace	Phthirusa abdita S. Moore	1	0	0	0	0	0	0	0
Pata-de-	Fabaceae	Bauhinia longifolia (Bong.)	1	0	0	0	0	0	1	1
Pé-de-	Euphorbiac	Croton goyazensis	1	0	0	0	0	0	1	1
Calunga	Simarouba	Simaba trichilioides A. St.-	1	0	0	0	0	0	1	1
Sete	Lythraceae	Cuphea mesostemon	1	1	1	0	0	0	1	3
Pau-doce-	Vochysiace	Vochysia tucanorum Mart.	1	15	8	0	0	0	1	24
Calção-de-	Boraginace	Cordia insignis Cham.	1	0	0	0	0	0	2	2
Douradinha	Rubiaceae	Palicourea xanthophylla	1	0	0	0	0	0	2	2
Salsa-	Liliaceae	Herreria salsaparrilha Mart.	1	1	0	0	0	0	2	3
Ipê-de-	Bignoniace	Stenolobium stans Seem.	1	5	0	0	0	0	2	7
Taiuiá	Cucurbitac	Siolmatra brasiliensis	1	2	1	0	0	0	2	5
Carobinha	Bignoniace	Memora nodosa (Silva	1	6	4	0	0	0	2	12
Canela de Ema	Velloziacea e	Vellozia flavicans Mart.		18	9	0	0	0	2	29
Araçá	Myrtaceae	Psidium firmum O. Berg.	1	0	0	0	0	0	3	3
Fruta-de-morcego	Fabaceae	Andira cuiabensis (Benth.) Benth.	1	1	0	0	0	0	3	4
Capotão	Vochysiace	Salvertia convallariodora A.	1	2	0	0	0	0	3	5
Pau-doce	Vochysiace	Vochysia rufa Mart.	1	3	0	0	0	0	4	7
Cansação	Euphorbiac	Jatropha urens L.	1	3	1	0	0	0	4	8
Cipó quina	Menisperm	Cissampelos ovalifolia DC.	1	2	1	0	0	0	5	8
Sangra	Euphorbiac	Croton salutaris M. Arg.	1	4	1	0	0	0	5	10
Pata-de-	Fabaceae	Bauhinia nitida Benth.	9	0	0	0	0	0	6	6
Azedinha	Oxalidacea	Oxalis hirsutissima Mart. ex	1	1	1	0	0	0	6	8
Tripa-de-	Fabaceae	Bauhinia glabra Jacq.	1	3	1	0	0	0	6	10
Ariticum,	Anonaceae	Annona dioica A.St.-Hil.	2	6	3	0	0	0	6	15
Cipó-mil-	Aristolochia	Aristolochia esperanzae	1	12	5	0	0	0	6	23
Guarantã	Olacaceae	Minquartia guianensis Aubl.	1	39	11	0	0	0	6	56
Genciana,	Fabaceae	Acosmium subelegans	3	7	0	0	0	0	7	14
Siputá	Hippocrate	Salacia elliptica (Mart. Ex	1	2	1	2	1	0	8	14
Grass	Poaceae	Digitaria insularis (L.)	1	49	30	0	0	0	10	89
Cana do	Costaceae	Costus arabicus L.	2	4	2	0	0	0	11	17
Tarumã	Verbenace	Vitex cymosa Bertero ex	2	9	5	0	0	0	13	27

Common name	Botanical family	Scientific name	Number of sources	M1="nc"	M2="nc"+"a"	M3="nc"+"d"	M4="nc"+"a"+d	F2=M5	F1=M6	Texts
Amargoso,	Apocynace	Aspidosperma polyneuron	4	60	25	0	0	0	14	99
No-de-	Malpighiac	Heteropterys aphrodisiaca	3	25	15	0	0	0	15	55
Calunga,	Simarouba	Simaba ferruginea A.St.-	11	8	2	1	1	0	15	27
Pata-de-vaca	Fabaceae	Bauhinia rufa (Bong.) Steud.	10	7	3	0	0	0	16	26
Carobinha	Bignoniace	Jacaranda decurrens	2	39	21	3	3	0	16	82
Taioba	Araceae	Xanthosoma violaceum	1	49	7	0	0	0	17	73
Purple IPÊ	Bignoniace	Tabebuia heptaphylla	2	33	16	2	1	0	17	69
Pata-de-vaca	Fabaceae	Bauhinia ungulata L.	1	21	11	2	2	0	17	53
Batata-de-teiú, purga	Euphorbiaceae	Jatropha elliptica (Pohl) Okem.	1	19	12	0	0	0	19	50
Scarlet	Solanaceae	Solanum gilo Raddi	1	93	43	2	1	0	20	159
Pata-de-	Fabaceae	Bauhinia splendens Kunth.	1	5	2	1	1	0	21	30
Batata-de-taiuiá,	Cucurbitaceae	Cayaponia tayuya (Vell.) Cogn.	1	17	12	0	0	0	22	51
Morcequeir	Fabaceae	Andira inermis (Sw.) Kunth.	1	32	13	0	0	0	23	68
Lixa, lixeira,	Dilleniacea	Curatella americana L.	2	44	11	4	1	0	27	87
Blackberry	Moraceae	Maclura tinctoria (L.) D.	2	23	11	0	0	0	32	66

Table 1. Displayed texts (No.) for 135 species used in diabetes control in Mato Grosso (Rieder, 2009) and their ordering by magnitude, according to searches and filling instructions in the Academic GA-Google search box [F1(insert direct from the 3 keywords: *"scientific name" "diabetes"* abstract")=M_6; F2(*"all in the title":"scientific name" "diabetes"*)=M_5; F3(M_1-*"in the title":"scientific name"*; M_2-*"in the title":"scientific name" "abstract"*; M_3-*"in the title":"scientific name" "diabetes"*; M_4-*"in the title":"scientific name" "diabetes" "abstract"*)], Cáceres, Mato Grosso(MT), Brazil, between December 21 and 29, 2012.

Common name	Botanical family	Scientific name	Number of sources	F3("in the title":"nc")				F2=M5 ("all in the title": "nc" "d")	F1=M6 ("nc" "d" "a")	Texts
				M1="nc"	M2="nc"+"a"	M3="nc"+"d"	M4="nc"+"a"+d			
Sucupira	Fabaceae	Pterodon pubescens	3	73	40	0	0	0	32	145
Negramina	Monimiace	Siparuna guianensis Aubl.	1	30	18	0	0	0	33	81
Caju-do-	Anacardiac	Anacardium humile A. St.-	13	43	26	3	3	0	33	108
Algodão-do-	Moraceae	Brosimum gaudichaudii	1	55	21	0	0	0	35	111
Chapéu-de-	Alismatace	Echinodorus macrophyllus	3	13	11	1	1	0	35	61
Quebra-	Euphorbiac	Phyllanthus tenellus Roxb.	1	30	15	6	3	0	35	89
Cordão-de-	Lamiaceae	Leonotis nepetifolia (L.)	2	12	5	0	0	0	38	55
Quina-do-	Loganiacea	Strychnos pseudoquina	3	15	9	3	2	0	38	67
Pata-de-	Fabaceae	Bauhinia cheilantha (Bong.)	1	14	10	3	3	0	38	68
Amargoso,	Fabaceae	Vatairea macrocarpa	4	21	10	3	3	0	38	75
Espinheira	Celastrace	Maytenus aquifolium	1	53	34	2	2	0	39	130
Carrapichin	Asteraceae	Acanthospermum australe	1	33	15	0	0	0	43	91
Sabugueiro	Caprifoliace	Sambucus australis Chan.	2	19	11	1	1	0	43	75
Sucupira-	Fabaceae	Pterodon emarginatus	2	62	37	11	8	0	52	170
Sangra d'Água	Euphorbiaceae	Croton urucurana Baill.	1	63	31	2	1	0	53	150
Stingray vine, grape, vegetable insulin	Vitaceae	Cissus verticillata (L.) Nicholson & C.E. Jarvis	7	17	11	10	9	0	55	102
Lemon	Rutaceae	Citrus limonum Risso	2	49	22	0	0	0	56	127
jackfruit	Moraceae	Artocarpus integrifolia L. F.	1	106	58	0	0	0	56	220
Cargueia-	Asteraceae	Baccharis articulata (Lam.)	2	30	15	6	6	0	57	114
Terramicina	Amarantha	Alternanthera brasiliana (L.)	1	42	21	1	1	0	58	123

Common name	Family	Scientific name							
Mine tea, water lettuce	Araceae	Pistia stratiotes L.	1	521	243	4	4	0	59	831
Sucupira	Fabaceae	Bowdichia virgilioides	8	60	36	8	8	0	59	171
Embuba	Cecropiace	Cecropia pachystachya	7	60	31	5	5	0	61	162
Seriguela	Anacardiac	Spondias purpurea L.	1	143	47	2	2	0	63	257
Mangaba	Apocynace	Hancornia speciosa Gomes	5	220	104	12	11	0	69	416
Carrapicho	Asteraceae	Acanthospermum hispidum	1	97	48	0	0	0	70	215
Jurubeba	Solanaceae	Solanum paniculatum L.	3	60	20	1	1	0	70	152
Guanandi	Clusiaceae	Calophyllum brasiliense	2	144	68	7	5	0	71	295
Genipapo	Rubiaceae	Genipa americana L.	5	157	61	1	0	0	73	292
Anador	Acanthacea	Justicia pectoralis Jacq.	1	90	36	4	2	0	82	214
Coentro-do-	Apiaceae	Eryngium foetidum L.	1	107	47	7	5	0	94	260
Wolf fruit, Lobeira	Solanaceae	Solanum lycocarpum St. Hil.	5	183	103	38	28	3	94	449
Carqueja	Asteraceae	Baccharis genistelloides	5	27	12	5	4	0	96	144
Jucá, pau-	Fabaceae	Caesalpinia ferrea Mart.	3	104	63	16	14	1	98	296
basking tea	Flacourtiaceae	Casearia sylvestris Sw.	1	140	74	5	3	0	110	332
Fedegoso	Fabaceae	Senna occidentalis (L.) Link	4	108	63	1	1	0	112	285
quinine	Simarouba	Quassia amara L.	1	203	64	9	8	0	119	403
field rosemary	Asteraceae	Baccharis dracunculifolia DC.	1	150	100	14	13	0	124	401
breadfruit	Moraceae	Artocarpus altilis (Parkison)	2	223	96	22	15	0	129	485
Marcela	Asteraceae	Achyrocline satureioides	1	162	87	6	5	0	148	408
Cancerosa	Celastrace	Maytenus ilicifolia (Schrad.)	2	280	131	11	9	0	175	606
Ipê-de-jardim	Bignoniaceae	Tecoma stans		212	108	40	30	2	191	583
Cress	Brassicace	Nasturtium officinale W.T.	1	194	100	7	2	0	221	524
Carqueja	Asteraceae	Baccharis trimera (Less.)	13	153	74	37	33	0	225	522
Hydrangea	Hydrangea	Hydrangea macrophylla	1	348	160	3	2	0	231	744
Comfrey	Boraginaceae	Symphytum officinale L.	6	239	81	7	6	0	242	575
Star fruit	Oxalidacea	Averrhoa carambola L.	25	510	237	36	23	3	246	105
Blackberry	Moraceae	Morus nigra L.	3	192	96	28	22	0	265	603
Pitanga	Myrtaceae	Eugenia uniflora L.	5	312	144	31	17	0	268	772
Pega-pinto	Nyctaginac	Boerhavia diffusa L.	1	51	31	13	12	2	289	398
Soursop	Anonaceae	Annona muricata L.	9	923	298	29	19	3	316	158
Pata-de-vaca	Fabaceae	Bauhinia forficata Link.	11	145	59	44	28	7	335	618
Annatto	Bixaceae	Bixa orellana L.	1	799	252	23	15	0	353	144
Strand	Asteraceae	Artemisia absinthium L.	1	338	139	6	4	0	358	845
Passion	Passiflorac	Passiflora edulis Sims.	2	176	565	41	32	3	387	278
Saffron	Zingiberace	Crocus sativus L.	2	120	587	32	21	1	462	230
Picão	Asteraceae	Bidens pilosa L.	6	606	298	44	32	7	471	145
Artichoke	Asteraceae	Cynara scolymus L.	2	103	398	37	20	4	475	196
Eucalyptus	Myrtaceae	Eucalyptus globulus Labill	3	350	139	27	16	0	480	541
Vassourinha	Scrophulariaceae	Scoparia dulcis L.	1	171	100	59	50	4	506	890
Eggplant	Solanaceae	Solanum melongena L.	1	284	117	44	35	1	528	461
Coffee	Rubiaceae	Coffea arabica L.	2	827	201	13	11	1	532	108
Andu beans	Fabaceae	Cajanus cajan (L.) Millsp.	1	3100	1310	39	34	0	561	5044
Bahia coconut	Arecaceae	Cocos nucífera L.	1	75	6	0	0	0	563	644
stone breaker	Euphorbiaceae	Phyllanthus niruri L.	5	416	250	62	47	0	663	1438
Elderberry	Caprifoliaceae	Sambucus nigra L.	2	525	235	19	13	3	687	148
Cashew	Anacardiac	Anacardium occidentale L.	15	163	629	67	55	3	711	309

Continuation of table 1

Common name	Botanical family	Scientific name	Number of sources	F3("in the title":"nc") M1="nc"	M2="nc"+"a"	M3="nc"+"d"	M4="nc"+"a"+d	F2=M5 (all in the title)	F1=M6 ("nc"d"a")	Texts
Stevia	Asteraceae	Stevia rebaudiana (Bertoni)	2	1480	584	182	131	5	722	3104
Tamarind	Fabaceae	Tamarindus indica L.	2	726	346	30	23	1	731	1857
Orange	Rutaceae	Citrus aurantium L.	3	688	279	37	29	0	768	1801
Basil	Lamiaceae	Ocimum basilicum L.	1	1460	780	36	25	1	860	3162
Pumpkin	Cucurbitac	Cucurbita pepo L.	1	2380	823	18	9	2	921	4153
Jambolão plum	Myrtaceae	Syzygium cumini (L.) Skeels	16	349	237	164	140	13	972	1875
Rosemary	Lamiaceae	Rosmarinus officinalis L.	1	1140	602	45	35	0	1010	2832
Green	Brassicace	Brassica oleracea L.	2	6900	275	51	40	1	1180	1092
Papaya	Caricaceae	Carica papaya L.	5	4110	154	40	32	1	1190	6913
Neem	Meliaceae	Azadirachta indica A. Juss	1	3500	172	146	104	18	1790	7278
Rice	Poaceae	Oryza sativa L.	1	2910	989	48	36	0	1880	4095
Soia	Fabaceae	Glycine max (L.) Merr.	1	1920	603	75	62	4	1950	2732
Onion	Liliaceae	Allium cepa L.	1	7400	262	98	68	4	2050	1224
Corn	Poaceae	Zea mays L.	1	4970	111	46	33	2	2630	6351
Saffron	Zingiberace	Curcuma longa L.	1	1670	897	114	78	8	2830	5597
Melon: Melão-de-São-	Cucurbitaceae	Momordica charantia L.	5	2010	965	527	364	62	3250	7178
Garlic	Liliaceae	Allium sativum L.	2	4910	181	180	135	9	3270	1031
Aloe	Liliaceae	Aloe vera (L.) Burm. F.	1	2710	1340	380	242	24	3640	8336
Sum in columns			370	1736	572	3252	2362	203	45475	2822

Continuation of table 1

Search procedures adopted with variants when filling out the search box (Search modes -$M_{1,2,3,4,5,6}$)	Simple absolute frequency: number (%) of species					Line Total
	Number of Classes of Displayed Texts ($C_{0,1,2,3,4}$)					
	A [absence]	B [one]	C [2 up to 10]	D [11 up to 50]	E [over 50]	
M_1= ["in the title":"scientific name"]	12(8,88)	4(2,96)	17(12,59)	30(22,22)	72(53,33)	135(100)
M_2= ["in the title":"scientific name"+"abstract"]	18(13,33)	8(5,92)	16(11,85)	37(27,40)	56(41,48)	135(100)
M_3= ["in the title":"scientific name"+"diabetes"]	51(37,77)	8(5,92)	30(22,22)	33(24,44)	13(9,62)	135(100)
M_4= ["in the title":"scientific name"+"abstract"+"diabetes"]	52(38,51)	12(8,88)	28(20,74)	33(24,44)	10(7,40)	135(100)
M_5= ["all in the title":"scientific name"+"diabetes"]	104(77,03)	8(5,92)	19(14,07)	3(2,22)	1(0,74)	135(100)
M_6= ["scientific name"+"abstract"+"diabetes"]	5(3,70)	5(3,70)	23(17,03)	30(22,22)	72(53,33)	135(100)
total – column	242	45	133	166	224	(135*6)=810
Note: Values and percentages (%) expected in the classes of number of texts of the 6 types of searches in view of the null hypothesis (H_0)	40,33(29,87)	7,5(5,55)	22,16(16,41)	27,66(20,49)	37,33(27,65)	135(100)

Table 2 - Number of plant species in five classes of text numbers displayed among 135 plants used to control diabetes in Mato Grosso (Rieder, 2009), by type of filling(p) in the Google Scholar search box and the search procedures adopted [M_6= F_1("*scientific name*" "*diabetes*" abstract"); M_5=F_2("*all in the title*":"*scientific name*" "*diabetes*"); F_3(M_1-"*in the title*":"*scientific name*"; M_2-"*in the title*":"*scientific name*" "*abstract*"; M_3- "*in the title*":"*scientific name*" "*diabetes*"; M_4-"*in the title*":"*scientific name*" "*diabetes*" "*abstract*")], Cáceres, Mato Grosso(MT), Brazil, between December 21 and 29, 2012.

Statistics	M₁= (nc in the title)	M₂=(nc in title;+a not in title)	M₃=(nc in the title; d not in the title)	M₄=(nc in the title; a and d not in the title)	M₅=(all in the title: nc and d)	M₆=(nc, a, d anywhere in the text)
N	135	135	135	135	135	135
Average	1286,378	424,3185	24,08889	17,4963	1,503704	336,8519
Standard error of average	455,5941	122,8184	5,442686	3,783206	0,526666	58,55333
Median	60	31	3	2	0	56
Average	0	0	0	0	0	2
Standard deviation	5293,525	1427,021	63,23829	43,95688	6,119304	680,3283
Coefficient of variation (%)	411,5062	336,3089	262,5206	251,2353	406,9487	201,9666
Variance	28021406	2036388	3999,082	1932,207	37,44588	462846,6
Symmetry	7,220951	5,919965	5,542773	5,234853	7,92048	3,113331
Symmetry standard error	0,208526	0,208526	0,208526	0,208526	0,208526	0,208526
Kurtosis	58,5613	38,48033	36,9314	33,78394	73,45934	10,02598
Kurtosis standard error	0,414153	0,414153	0,414153	0,414153	0,414153	0,414153
full range	49700	11100	527	364	62	3640
Smallest value	0	0	0	0	0	0
Highest value	49700	11100	527	364	62	3640
Sum	173661	57283	3252	2362	203	45475
Percentile 10	1	0	0	0	0	2
Percentile 50	60	31	3	2	0	56
Percentile 75	348	160	28	17	0	335

Table 3 - Statistics regarding the amounts of texts displayed for 135 species used to control diabetes in Mato Grosso (Rieder, 2009), given the search and fill-in resources adopted in the Academic Google search box [F1 (direct insertion of the 3 key words: **nc** *"scientific name"*; **d** *"diabetes"* **a** abstract" M6: *"nc" "d" "a"*); F2(*"all in the title":"nc" "d"*)=M5; F3(M1- *"in the title":"nc"*; M2- *"in the title":"nc" "a"*; M3- *"in the title":"nc" "d"*; M4- *"in the title":"nc" "d" "a"*)], Cáceres (MT), Brasil, between December 21 and 29, 2012.

The lowest percentage of species included in classes A, B, C, D and E of the amount of retrieved texts occurs in M_6, M_1, M_1, M_5 and M_5, respectively. While the highest percentage occurs in the modalities M_5, M_4, M_3, M_2 and, M_1 and M_6, respectively.

TEXT QUANTITY CLASSES (A_T) WITH LOWEST AND HIGHEST % OF SPECIES IN M_T

The lowest(-) and the highest(+) percentage of species in the search modalities ($M_{1...6}$) were registered in the following classes: M_1:[-B(2,96%) and +E(53,33%)]; M_2:[-B(5,92%) and +E(41,48%)]; M_3:[-B(5,92%) and +A(37,77%)]; M_4:[-E(7,40%) and +A(38,51%)]; M_5:[-E(0,74%) and +A(77,03%)]; M_6:[-A and B(3,70%) and +E(53,33%)]

Less discriminatory modalities retrieve more documents for a greater number of species and less for a few or absence of texts. While the inversion of this distribution of cases to the more discriminatory ones is established.

SPECIES WITH STUDIES FOR DIABETES ANNOUNCED IN THE TITLE (M_5)

Only 31 out of 135 species exhibited documents with diabetes studies advertised in the title. These 31 species are also in the list of the 54 best placed (in number of texts) in the results of the search mode:M_6.

In M_5 there were eight species with only one document, nineteen having the range from 2 to 10, three with 11 to 50 and, with more than 50 documents, only the melon: Melão-de-São-Caetano.

Six of the ten most cited in use for diabetes in TM did not present in the search (M_5) diabetes-specific study.

This reveals that there is a lack of studies and/or availability of research results for plants that are used with high frequency to control diabetes in this region of the world. (Mato Grosso, Brasil).

RETRIEVAL RANGE AND TEXT FILTERING POWER IN SEARCHES

Text recovery for *"scientific name"* without and with filtering for *"abstract"*

In modality: M_2 filtered M1 documents for 121 species. Both modalities in twelve species did not recover texts (absence) and, in two, only one for each species.

In M_4 also filtered out M_3 for 66 species, but for 51 there was an absence of texts for both modalities; while for 7, 3, 4, 1, 1, 1 and 1 species there were, respectively, 1, 2, 3, 4, 5, 6 and 8 texts so much for M_4 clike to M_3.

Therefore, in M_1 and in the M_3 more texts were retrieved, but without qualifying them as to whether they were academic or not. To select academic documents about the species referred to in the title, the appropriate modalities, respectively, are M_2 and M_4. These reveal documents for the species in question, in the case of M_2 and, for species relating it to diabetes, in the case of M_4.

M2 and M4 modalities, by using the keyword *"abstract"* retrieve fewer texts for species than pairs M1 and M3 that do not apply that keyword. Thus M_2 and M_4 filter more, discarding non-academic texts.

The retrieval of texts for the species of interest, launched in the search box, grows in quantity from the modality:M_5, in the following order: $M_5<M_4<M_3<M_6<M_2<M_1$.

To retrieve texts that deal with the species in relation to diabetes, the modalities: M_1 and M_2 are not applied. In these, the capture instruction does not include the keyword *"diabetes"*(*"d"*). Those that include *"d"*, the number of documents selected, expressed in the following order: $M_5<M_4<M_3<M_6$. But the texts retrieved by M_5 are the most relevant.

RECOVERY RANGES

The ability to recover is in the following order: $M_1 > M_2 > M_6 > M_3 > M_4 > M_5$. The first two explore the *"nc"* in the title and without adding "d"; the third instructs the search with the three keywords (*"nc"*, *"d"*, *"a"*), but anywhere in the document (it doesn't need to be in the title); the fourth and fifth explore the *"nc"* in the title and add *"d"* in the search box instruction; the last instructs that so much *"nc"* like *"d"* are in the title. This demonstrates that the most focused and rigorous filtering takes place in M_5 given the variations in search instructions in each applied modality.

RECOVERY AND FILTERING POWER

The following proportion and inequality between recovery modalities occurred: $M_1 > M_2 (= 0,329 M_1) > M_6 (= 0,793 M_2) > M_3 (= 0,071 M_6) > M_4 (= 0,726 M_3) > M_5 (= 0,085 M_4 = 0,004 M_6)$.

The filtering power of texts related to diabetes of plant species was established in the following relationship between search modalities:

(M6-M6)/M6; (M6-M5)/M5; (M5-M4)/M4; $M_5 > M_4 > M_3 > M_6$;

$M_6 < M_3 (= 12,983 M_6) < M_4 (= 0,376 M_3) < M_5 (= 10,635 M_4 = 15,019 M_3 = 223,014 M_6)$;

$M_5 (= 0,9955 M_6 = 0,9375 M_3 = 0,9140 M_4) > M_4 (= 0,9480 M_6 = 0,2736 M_3) > M_3 (= 0,9284 M_6) > M_6$.

For 46 species there are documents recovered by M_6, but in M_4 are absent.

The format *"all in the title"* (F2=M_5) filters the recovery of texts more than the formato *"in the title"* (F3=M_1;M_2;M_3;M_4) and, this is more than the format you look for anywhere in the document. (F1=M_6). Thus, the three formats have the following order of filtering power: F2>F3>F1. Additionally, this power is also governed by the quality and quantity of keywords entered in the search instruction (completion of the search box). The more keywords that are applied, the fewer documents will be retrieved and the more selective the search. The application of "in quotation marks" to keywords is also added to the filtering refinement. "Enclosed" keywords retrieve fewer documents than unquoted ones.

The more specific the search instruction, the greater the document filtering.

However, it is suggested to continue investments in the development of more efficient and effective tools in the control and recovery of data, information and documents of general and specific interests, as well as in the training of instructor agents, especially in user support services, as well as provide training opportunities for users of the academy, whether students or professors and researchers. Because the option of online searches from secondary academic sources, available on the internet, is only going to grow. These observations are supported by several scholars. (CUENCA *et al.*,2003; PINHEIRO *et al.*, 2008; NARUKAWA *et al.*, 2009; PONTES JUNIOR and TÁLAMO, 2009; MORAES e OLIVEIRA, 2010).

FINAL CONSIDERATIONS

Databases and virtual libraries were accessed via the internet, between December 21-29, 2012, through the search engine Google Scholar (GA). The procedures adopted retrieved scientific documents about plant species used to control diabetes, displayed on the screen of the computer used.

Among the plant species used for diabetes in Mato Grosso, some are already well studied, others with studies that are not yet conclusive and a considerable number of species without any scientific study that can validate or not the popular application.

In the four searches you enter in the search box ""*in the title*""(M3,M4,) or ""*all in the title*""(M5) and the keywords "*scientific name*" and "diabetes", the melon: Melão-de-São-Caetano(*Mormodica charantia*) is the species

with the highest number of retrieved texts.

Aloe vera (Aloe vera) reveals more retrieved texts when not used in GA search box, resources *in the title*, or *all in the title*, mut directly the keywords *scientific name*,*diabetes* and *abstract* (M_6). This species is the 2nd that more texts exhibited in the previous procedures (M_3, M_4 and M_5);

Unequal proportionality in the ability to recover modalities is as follows: M_1 > $M_2(=0,3298M_1)$ > $M_6(=0,7938M_2)$ > $M_3(=0,0715M_6)$ > $M_4(=0,7263M_3)$ > $M_5(=0,0859M_4=0,00446399M_6)$;

While filtering power is in opposition to resilience, in the following proportions of inequality: M_6 < $M_3(=12,9837M_6)$ < $M_4(=0,3767M_3)$ < $M_5(=10,6354M_4=15,0197M_3=223,0147M_6)$ or $M_5(=0,9955M_6=0,9375M_3=0,9140M_4)$ > $M_4(=0,9480M_6=0,2736M_3)$ > $M_3(=0,9284M_6)$ > M_6

The more you specify the search instruction, the more filtering you will have in the retrieval.

The search procedures to retrieve documents *online* can be applied to any thematic focus of the different areas of knowledge, including interdisciplinary ones, but they need technical updates and with their users properly instructed.

REFERENCES

BIOLCHINI, J. C. de A.; GIORDANO, R. Comportamento informacional na produção científica: busca e recuperação da informação na construção do campo da Ciência da Informação. 2011. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12, 23-26 out. 2011, Brasília. **Anais**…. Brasília, 2011. p. 944-963.Available on the website:< http://ridi.ibict.br/handle/123456789/90>. Accessed on: March 26, 2014.

BONFIM, M. E. Recuperação de documentos-texto usando modelos probabilísticos estendidos. **Iniciação Científica CESUMAR**. Maringá, v. 11, n. 2, p. 149-157, Jul./Dez. 2009. ISSN 1518-1243. Available on the website:<http://www.cesumar.br/pesquisa/periodicos/index.php/iccesumar/article/viewFile/1282/884>. Accessed on: March 19, 2014.

CAMARA, M. C. C.; COSTA, L.; MARINHO, C. L. C.; GUILAN, M. C. R. A produção científica sobre intoxicações por agrotóxicos na região serrana do Rio de Janeiro. **O Mundo da Saúde**, São Paulo, 32(3), p. 268-274, jul.-set. 2008. Available on the website:<http://www.saocamilo-sp.br/pdf/mundo_saude/63/268-274.pdf>. Accessed on: August 25, 2013.

CAREGNATO, S. Google Acadêmico como ferramenta para os estudos de citações: Avaliação da precisão das buscas por Autor. **PontodeAcesso**, América do Norte, 5(3), fev. 2012. Available on the website:<http://www.portalseer.ufba.br/index.php/revistaici/article/view/5682>. Acesso em: 09 Mar. 2014.

COELHO, C. **Um Repositório Digital para a U. Porto: Relatório Preliminar.** U. Porto. Biblioteca Virtual Disponível <http://sigarra.up.pt/up_uk/WEB_GESSI_DOCS.download_file?p_name=F1368788598/repositorio-vpreliminar.pdf>. Access at:. March 3, 2014.

CORRÊA, A. C. G. Recuperação de documentos baseada em informação semântica no ambiente AMMO. São Carlos: UFSCar, 2005. 92 p. Dissertação (Mestrado, Programa de Pós- Graduação em Ciência da Computação, Universidade Federal de São Carlos UFSCar) São Carlos, Brasil, 2005). Available on the website: <http://www.bdtd.ufscar.br/htdocs/tedeSimplificado//tde_busca/arquivo.php?codArquivo=485>.Accessed: March 21, 2014

CUENCA, A. M. B.; ALVAREZ, M. do C. A.; FERRAZ,M. L. E. de F.; GARBIN, V. L. O usuário remoto de uma biblioteca acadêmica na área de saúde pública. **Informação. & Sociedade: Estudos**, João Pessoa, v. 13, n. 2, p. 91-109, jul./dez. 2003. Available on the website: <http://www.ies.ufpb.br/ojs2/index.php/ies/article/view/92/1563>. Accessed on: June 3, 2014.

DELICATO, F. C.; PIRMEZ, L.; DA COSTA CARMO, L. F. R. Multiagentes para a filtragem de páginas Web. In: Simpósio Brasileiro de Redes de Computadores, 19., 2001, Florianópolis, Santa Catarina. **Anais**… Florianópolis, SBRC, 2001. p. 512-526. Available on the website:<http://ce-resd.facom.ufms.br/sbrc/2001/033.pdf>. Accessed on:: March 23, 2014.

FERNANDES, J. C.; PINTO, V. B.; AZEVEDO NETO, C. X. de. Representação indexal na WEB: estudo do sintagma "História da Paraíba" nos sites Alta Vista e Google. Informação & Sociedade: Estudos, João Pessoa, v. 15, n. 2, p. 181-212, jul./dez. 2005. Available on the website: <http://www.ies.ufpb.br/ojs2/index.php/ies/article/view/37>. Accessed: June 3, 2014.

FERNANDES, R. F.; TEZA, P.; MIGUEZ, V. B.; DANDOLINI, G. A.; SOUZA, J. A. de; ABREU, A. F. de. Ferramentas *Scup* e *Many Eyes* na identificação de oportunidades para o *Front End* da inovação**. DataGramaZero - Revista de Informação**, v.14, n.5, p.1-18, out 2013. Available on the website: <http://www.dgz.org.br/out13/Art_01.htm>. Accessed on: March 3, 2014.

GOOGLE. **Google Acadêmico**: Sobre o Google Acadêmico. 2014. <Available on the website: <http://scholar.google.com.br/ intl/pt-BR/scholar/about.html>. Accessed on: March 23, 2014.

IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia. Página inicial. Informação para a Pesquisa. **Repositórios Digitais**. Available on the website:<http://www.ibict.br/informacao-para-ciencia-tecnologia-e-inovacao%20/repositorios-digitais>. Accessed on: March 2, 2014.

LAMAS, W. de Q. Como realizar pesquisa científica pela Internet em periódicos indexados por meio do portal de periódicos da CAPES. **Rev. Ciênc. Exatas**, Taubaté, v.9/10, n.1-2, p.83-87, 2003/2004. Available on the website: <http://periodicos.unitau.br/ ojs-2.2/index.php/exatas/article/viewFile/332/523>. Accessed on: March 23, 2014.

LEMOS, C. J. de F. **Arquitetura da Informação aplicada a websites de Governo Eletrônico**: estudo de caso do Tribunal de Justiça da Bahia. Salvador, 2009. 176 p. Dissertação (Mestrado – Instituto de Ciência da Informação) Universidade Federal da Bahia - UFBA. Available on the website: <https://repositorio.ufba.br/ri/bitstream/ri/7924/1/Lemos_C%20J%20F_Arquitetura%20 Informa%C3%A7%C3%A3o%20websites%20Governo%20Eletr%C3%B4ni.pdf >. Accessed on: March 2, 2014.

MACHADO DE MORAES, M. H.; DIAS MIRANDA, A. C. Produção do conhecimento sobre o sistema eletrônico de editoração de revistas (SEER) no Brasil nos anos de 2003 a 2010. *Encontros* **Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 16, n. 32, p. 27-40, 2011. ISSN 1518-2924. DOI: 10.5007/15182924.2011v16n32p27 Available on the website:<http://www.redalyc.org/src/inicio/ArtPdfRed.jsp?iCve=14720012003>. Accessed on: March 19, 2014.

MARTINS, D. A emergência da análise de redes sociais como campo de pesquisa: perspectiva da análise da produção científica em português e espanhol a partir do Google Academico.**Alexandría : Revista de Ciencias de la Información**, vol. 5, n. 8. 2011, [Journal article (On-line/Unpaginated)]. Available on the website:<http://eprints.rclis.org/16876>. Accessed on: May 11, 2013.

MEDEIROS, D. **Sistemas de recuperação e disseminação seletiva da informação em unidades de informação empresariais: quatro pilares e quatro atributos** Florianópolis (SC), 2008. 61 f. Trabalho de Conclusão de Curso (graduação) – Universidade do Estado de Santa Catarina, Florianópolis. (SC), 2008. Available on the website:<http://www.pergamum.udesc.br/dados-bu/000000/00000000000B/00000B8C.pdf>. Accessed on: March 23, 2014

MILANEZ, D. H. **Nanotecnologia: Indicadores tecnológicos sobre os avanços em materiais a partir da análise de documentos de patentes.** São Carlos: UFSCar, 2011. 176 f. Dissertação (Mestrado), Universidade Federal de São Carlos, 2011. Available on the website: <http://www.bdtd.ufscar.br/htdocs/tedeSimplificado//tde_busca/arquivo.php?codArquivo=4653>. Acesso em: 23 mar. 2014.

MORAES, A. F. de; OLIVEIRA, T.M de. Experiências relacionadas ao levantamento de teses e dissertações. **Informação & Sociedade: Estudos**, João Pessoa, v.20, n.1, p. 73-81, jan./abr. 2010 Available on the website: <http://www.ies.ufpb.br/ojs2/index. php/ies/article/view/4001/3421>. Accessed on: June 3, 2014.

MUGNAINI, R.; STREHL, L. Recuperação e impacto da produção científica na era Google: uma análise comparativa entre o Google acadêmico e a Web of Science. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, [S.l.], v13, nesp.1, p. 92-105, abr. 2008. ISSN 1518-2924. 10.5007/1518-2924. Available on the website: <https://periodicos.ufsc.br/ index.php/eb/article/view/1518-2924.2008v13nesp1p92>. Accessed on: March 9, 2014.

NARUKAWA, C. M.; LEIVA, IG; FUJITA, M.S.L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de Odontologia. Informação & Sociedade:Estudos, João Pessoa, v.19, n.2, p. 99-118, maio/ago. 2009. Available on the website: <http://www.ies.ufpb.br/ojs2/index.php/ies/article/view/2925/3028>. Accessed on: June 3, 2014.

NOGUEIRA, F. C. **Gêneros midiáticos, internet e contexto escolar**: relações entre fazeres e saberes. Cuiabá:UFMT, 2012. 132 f. Dissertação (mestrado) – Instituto de Educação, Programa de Pós-Graduação em Educação, Universidade Federal de Mato Grosso-UFMT. Available on the website: <www.ie.ufmt.br/ppge/dissertacoes/index.php?op=download&id=355>. Accessed on: March 9, 2014.

PAGANELLI, C. J. M. Responsabilidade objetiva dos provedores de conteúdo na internet. **Âmbito Jurídico**, Rio Grande, v.14, n. 95, dez. 2011. Available on the website: <http://www.ambitojuridico.com.br/site/?n_link=revista_artigos_leitura&artigo_id=10772&revista_caderno=7>. Accessed in March 2014.

PINHEIRO M. I. da S.; SILVA, E. R. P. da; MARQUES, N. J.;SILVA S. M. da. A utilização das fontes informacionais na rede particular de ensino médio de Rondonópolis – MT. **Informação. & Sociedade: Estudos**., João Pessoa, v.18, n.3, p. 143-150, set./dez. 2008. Available on the website:<http://www.ies.ufpb.br/ojs2/index.php/ies/article/view/1720/2697>. Accessed on: June 3, 2014.

PINTO, A. L.; PONTES JUNIOR, J. de. Índice científico da Ciência da Informação: Comparativo dos autores com maior visibilidade no ISI e no Google Acadêmico. **Informação & Informação**, [S.l.], v. 13, n. 1, set. 2008. ISSN 1981-8920. Available on the website: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/1801>. Accessed on: March 9, 2014. doi:10.5433/1981-8920.2008v13n1p.

PONTES JUNIOR J. de; TÁLAMO M. de F. G. M. Alfabetização digital: proposição de parâmetros metodológicos em competência informacional. **Informação & Sociedade:Estudos**, João Pessoa, v.19, n.2, p. 81-98, maio/ago. 2009. Available on the website:<http://www.ies.ufpb.br/ojs2/index.php/ies/article/view/2990>. Accessed on: June 3, 2014.

RAMOS, C.; MUNHOZ, D. P. A subjetividade da relevância na recuperação da informação: análise a partir de imagens representativas. **Biblos:** Revista do Instituto de Ciências Humanas e da Informação, v. 25, n.1, p.69-79, jan./jun. 2011. Available on the website:<http://www.seer.furg.br/index.php/biblos/article/view/1932/1229>. Accessed: March 23, 2014.

RIEDER, A. **Plantas medicinais utilizadas para o controle de diabetes em Mato Grosso, Brasil**. Cuiabá – MT: UFMT, 2009. 72p. Relatório de Estágio de Pós-doutorado (INSTITUTO DE BIOCIÊNCIAS - UNIVERSIDADE FEDERAL DE MATO GROSSO), Grupo de Pesquisas da Flora, Vegetação e Etnobotânica – FLOVET, 2009.

ROSA, R. **O potencial educativo das TICs no ensino superior: uma revisão sistemática.** Uberaba:UNIUBE 2009. 121 f. Dissertação (Mestrado em Educação). Universidade de Uberaba, Uberaba, MG, 2009. Available on the website:<http://www.uniube.br/biblioteca/novo/base/teses/BU000181088.pdf>. Accessed: March 23, 2014.

ROSSINI, T.; SANTOS, E. A mediação docente como interatividade: elementos essenciais para a educação na modalidade on-line em ambientes virtuais. **Informática na Educação:** teoria & prática. Porto Alegre, v. 16, n. 2, p. 187-202, jul./dez. 2013. Available on the website:<http://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/26405> Accessed on: February 28, 2014.

SALES, R. de; VIERA, A. F. G. Grupos e linhas de pesquisa sobre recuperação da informação no Brasil. **Biblios** [*On-line*] Perú: Julio Santillán Aladana Ed., (28), p. 1-14, abr.-jun 2007. Available on the website: <http://www.redalyc.org/articulo.oa?id=16114070004>.

SEIBEL JÚNIOR, H. **Recuperação de informações relevantes em documentos digitais baseada na resolução de anáforas**. Vitória: UFES, 2007. Dissertação (Mestrado em Informática), Centro Tecnológico, Universidade Federal do Espírito Santo. Available on the website:<http://hsjunior.files.wordpress.com/2012/10/dissertacao_hilario.pdf>. Accessed on: March 23, 2014.

SOUZA, F. A. S. de. **Recuperação da informação na Web: uma análise da ferramenta de busca Google Acadêmico**. 2009. 73 f. Monografia (Bacharelado) - Curso de Graduação em Biblioteconomia, Departamento de Biblioteconomia, UFRN, Natal, 2009. Available on the website:<http://repositorio.ufrn.br:8080/monografias/handle/1/260 Accessed: March 9, 2014

AUTHORIZATIONS / RECOGNITION

We, the undersigned authors, by submitting the work to Atena Editora and IJHS for the purposes of its publication, declare to be responsible for the content of this work generated by us and, we inform that there is no conflict of interest.

Cáceres(MT), November 8, 2021

Arno Rieder   Fabiana Aparecida Caldart Rodrigues