NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Using the Pagerank Algorithm to Explore a Physician Referral Network

Farid Nicolás Calderón Gomez

Dissertation report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2017

Title:

Subtitle:

Student
full name

MAA

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

USING THE PAGERANK ALGORITHM TO EXPLORE A PHYSICIAN REFERRAL NETWORK

by

Farid Nicolás Calderón Gomez

Dissertation report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Advisor / Co Advisor: Fernando Bação

November 2017

ABSTRACT

PageRank algorithm is nowadays the baseline in search engine being the tool behind Google's operation. However, since it was defined many different approaches have been tried in different knowledge areas. In this document a description of this tool, specification of the different areas where it has been used and new application of this algorithm in a physician referral network are presented.

KEYWORDS

Referral network; PageRank; Search engine

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Everybody agrees that the importance and influence of Google search engine in our everyday life are very significant. According to some of its statistics, it has reached 300 million active users monthly, processing over 40,000 search queries every second on average, which translates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide (internet live stats google search statistics). Obviously, its success is largely due to the quality of the algorithm used to obtain the results that users need, the PageRank algorithm.

The introduction of PageRank completely revolutionized the way search engines worked until that moment. Having a link structure approach to the web, and the ability to measure the relative importance of pages, based not only on its backlinks but also in the importance of those edges, PageRank is able to retrieve the most relevant results for each user search.

However, the PageRank algorithm didn't only change the way we search for information on the web, it has contributed to change many other areas. Examples of the application of the PageRank algorithms or slightly different versions come from fields such as: medicine [5], recommendation systems [4], sports team's rankings [6], the study of ecosystems [7], and others.

The use of the original PageRank algorithm in different contexts from what was its original objective stems from the fact that so many problems can be cast as the study of the behavior and relationships between nodes of a network. In fact, any problem in which a link structure can be constructed is a candidate for the application of the PageRank algorithm. The application of the algorithm will allow the identification of the most essential elements in the net. This identification has been found to be very useful in many different areas of science and technology.

In this paper, we will use the PageRank algorithm to assess the relative importance over 94 physician specialties in a Physician Referral Network and understand the main characteristics of the whole Physician Referral structure. The assessment of the different physician specialties is not only based on their amount of traffic (number of patients), but we also consider its connections with other physician specialties. We want to understand what are the most physician specialties, that have the highest potential to affect the overall flow of a patient in the health care company.

This paper is organized as follows, the next section is dedicated to reviewing the many different application of the PageRank algorithm. In the third section, we will delve into the details of PageRank, we identify the most important concepts, explain the algorithm in detail and show its relationship with Markov Chains. In the next we present the application to of the Page Rank algorithm to the study of the importance in a Physician Referral Network.

2. LITERATURE REVIEW

PageRank algorithm bases all its theory on the idea of seeing the web as a network between web pages. It is a method for computing a ranking for each web page based on the graph of the web. [1]. For this reason, any problem in any context that includes set of individuals connected between them building a link structure is a candidate for applying this algorithm.

The objective of this exercise is determining the most essential element of the network such as in google, where the most relevant pages are returned in the first places. Some of its applications are explained in this section.

2.1. GROUP RECOMMENDATION

A group recommendation network analyzes social and trust relationship between users. PageRank is used to determine the member's importance in the group, this importance is traduced in influence, which reflects the group social characteristics. This application appears as an answer for information overload to determine group leaders for decision making. [4]

In this situation, the network graph relates the way group members trust between them, Figure 1 shows for instance, individual 1 trust in individual 7 but individual 7 does not trust in him. Individuals 5 and 6 trusts between them in both directions.

The influence of a person depends not only in the number of people that trust in him but in how influential they are, following the idea of PageRank weights.

As an example of this, given a social group, a company can calculate the importance of each member of that group and based on the result decide how is going to apply some strategy for example spending money only in marketing for the most important and influential individuals knowing they will recommend its product with the rest of the group.

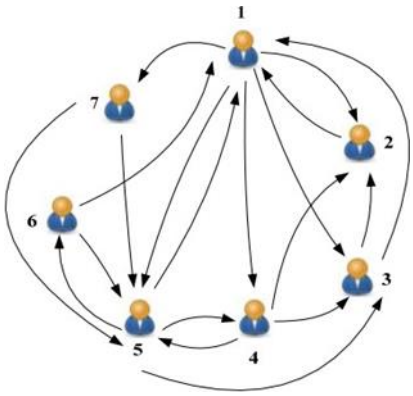

Figure 1 - Group Diagram. Source from [4]

2.2. MEDICINE

Metastasis is the spread of cancer cells to new areas of the body (often by way of the lymph system or bloodstream). A metastatic cancer, or metastatic tumor, is one which has spread from the primary site of origin (where it started) into different area(s) of the body.[1]

Cancer's treatment most of the times becomes difficult because of the mentioned process due to the lack of knowledge about which is the next point the cancer is going to attack. To solve this problem a network that provides the relationship between body organs as nodes and the way they are involve in cancer, can help to determine the body organ which have to be treated depending on the value achieve.

[5] recreates the explained exercise for Lung Cancer in which based on real cases a body organ's network is created, the results will be described in terms of a "random walker" leaving the lung and traversing the networks moving from site to site along one of the outgoing edges available to it at the site it is leaving, choosing a given edge with the probability corresponding to its weighting. [5]

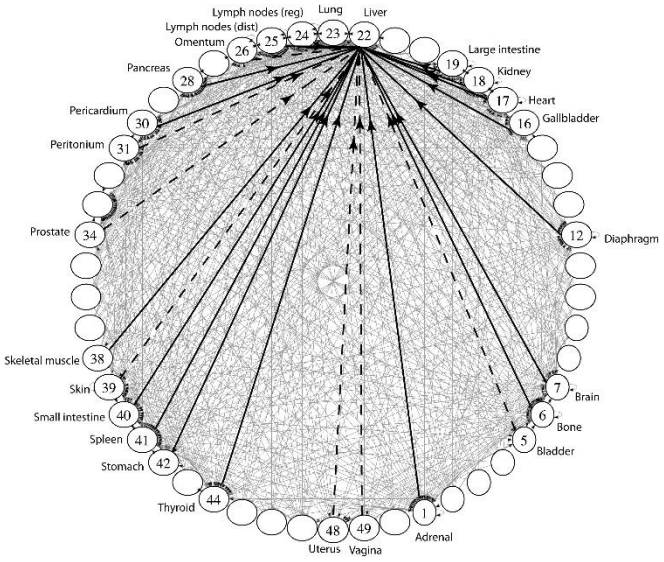

Figure 2 - Construction of Metastatic sites. Source from [5]

2.3. SPORTS TEAMS RANKING

Football is without doubt the most famous sport around the world. The FIFA World Cup gathers the countries fans each 4 years since 1930.

[6] creates a new rank for the soccer teams based on the Page Rank Algorithm. A network is created by considering the teams as the nodes and the links as the games between them. Different statistics such as matches won were considered for creating different metrics that assign weights to the links in the graph.

[1] Cancer center Webpage definition < http://www.cancercenter.com/terms/metastasis/ >

In figure 3 the results for all the countries involved is shown. After comparing the algorithm results with the current FIFA classification, the results are consistent. Based on the image, it can be said that 5 groups can be identified, thus throw this exercise PageRank algorithm demonstrates its ability as a possible clustering tool too.

It is important to remark that it is possible to apply this tool for different sports considering the results over history.

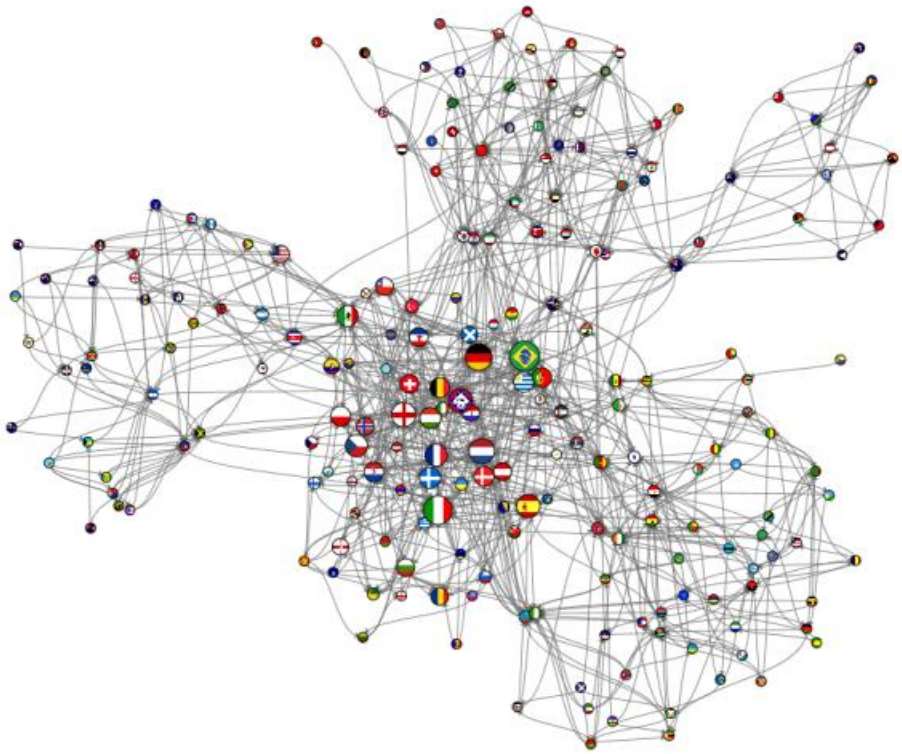

Figure 3 - World Teams Network Structure. Source from [6]

2.4. ECOSYSTEMS

Unfortunately, extinction of species has acquired a bigger probability in the last years because of how humans use the nature and the low importance given to the environment. Hence the interest in forecasting the impact of this phenomenon on ecosystems, which is not always easy due to species quantity.

Figure 4 shows a food web in which each specie is a node and nutrient flows are represented by edges. For this approach, the interpretation is different than the original interpretation of PageRank: the importance of an individual depends on the number of forward links it has: a specie is important if important species rely on it for their survival.

The proposed method identifies the species of maximum importance by determining which extinctions lead to the fastest ecosystem collapse. Moreover, the algorithm bridges the gap between qualitative (who eats whom) and quantitative (at what rate) descriptions of food webs [7].

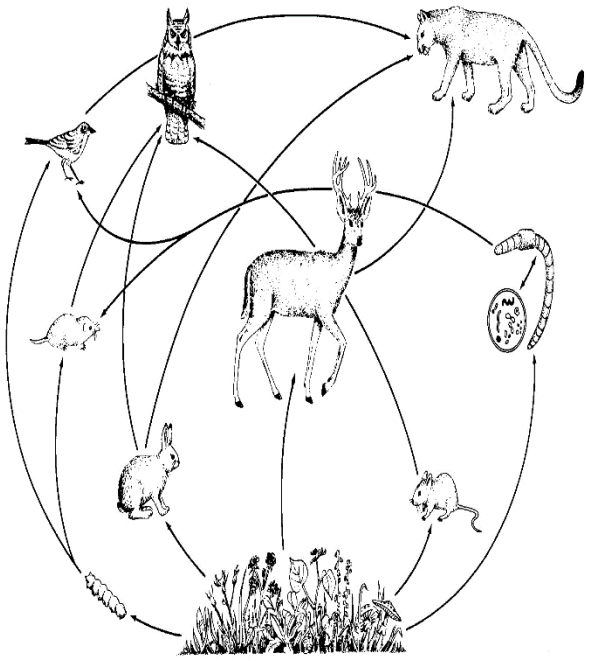

Figure 4 - Food Web

3. THEORETICAL REVIEW

PageRank is an attempt to see how good an approximation to “importance" can be obtained just from the link structure. In other words, a page has high rank if the sum of the ranks of its backlinks is high. This covers both the case when a page has many backlinks and when a page has a few highly ranked backlinks [1].

Figure 5 shows the way input links distribute their corresponding PageRank depending on the number of output links they have. In general, the weight (W_i) with which each page will affect the other pages relates the PageRank value and the number of forward links as follows:

$$W_i = PagerankValue_i * \left(\frac{1}{\#\ forwardlinks_i} \right) \quad (1)$$

In general given a web page *u*, the simple definition of PageRank in order to calculate the PageRank of page *u* (R(*u*))is given by:

$$R_u = c \sum_{u \in B_u} \frac{R(v)}{N_v} \quad (2)$$

Let B_u be the set of pages pointing u, R_v is the PageRank of the page v, and N_v represents the number of outlinks of page v.

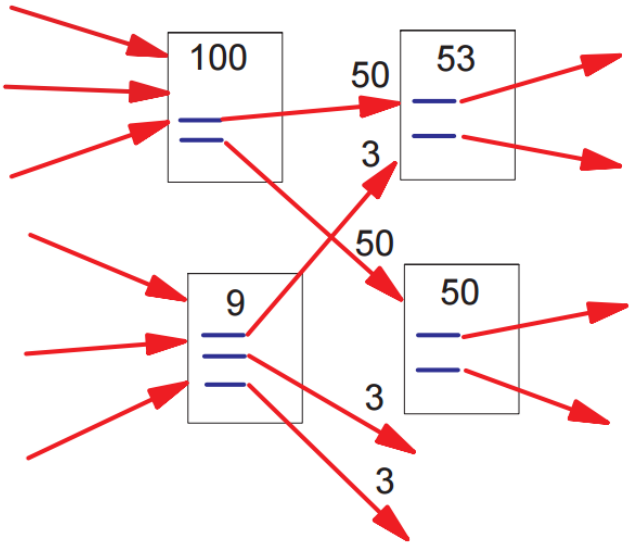

Figure 5 - Simplified PageRank Calculation [1]

Having this, the iterative process for obtaining the results is presented in [3]. Assuming page A has T1 ... TN which point to it. The parameter d is a damping factor which value can be set between 0 and 1. Also C(A) is defined as the number of links going out of page A. The PageRank of page A (PR(A)) is:

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(TN)}{C(TN)}\right) \quad (3)$$

This equation is recursive but it may be computed by starting with any set of ranks and iterating the computation until it converges. However, in this definition exists a problem when two pages are pointing between them, they do not point to another page and another page points one of them, because during the iteration rank is going to be accumulated during the loop but never distributed, thus the algorithm creates a kind of trap.

For resolving this problem google creators made a rank source, the new formal definition of the algorithm according to [1] is given by:

Let E(u) be some vector over the Web pages that corresponds to a source of rank. Then, the PageRank of a set of Web pages is an assignment, R'*, to the Web pages which satisfies:*

$$R'(u) = c\sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u) \quad (4)$$

such that c is maximized and $\| R' \|_1$ *=1*

PageRank algorithm is based on this idea, in which a random surfer is imagined navigating through the web as a random walker, he follows the edges in an uniformly way with a given probability α, but randomly he will jump to any page with probability $1-\alpha$. In other words, the surfer is going to keep clicking on links from a random page until a point when he gets bored and starts another random page [3].

Here, for each page, the probability of being visited by the surfer is precisely its PageRank, thus the pages where we can find the surfer must often are the important ones. The damping factor is the probability at each page the random surfer will get bored and request another random page.

Using this surfer model for a given web structure, PageRank obtains a probability distribution as it corresponds to the standing probability distribution of a random walk on the graph of the web [1].

The relationship between PageRank and Markov chains can be seen based on the Random Walker concept. The idea of having a surfer clicking on links starting from a random page with a probability of “jumping" to another random page in any time can be traduced as a Markov chain to predict the behavior of the surfer that travels from one state to another considering only the current solution.

As it was seing in Section 3 the number of links to and from a web page determines its importance. [9] defines the basic PageRank of a page p (PR(P)) as:

$$PR(p) = \sum_{q \in pa_p} \frac{PR_q}{|O_q|} \tag{5}$$

Where pa and O_q are the set of backlinks and forward links of p respectively.

The PageRank as was defined in Section 3 is a recursive algorithm. It assigns an initial value of $PR_p^0 = \frac{1}{n}$ being n the total number of pages on the web. Taking that into account equation (5) iterates as follows where $k = 0,1,2,\ldots$ is refer to the iteration in which the process is:

$$PR_p^{k+1} = \sum \frac{PR_q^k}{|O_q|}_{q \in pa_p} \tag{6}$$

(6) can be expressed as the following expression, where q^k is the PageRank vector at iteration k and T is the transition matrix for the web.

$$q^{k+1} = Tq^k \tag{7}$$

PageRank creators built an [8] irreducible aperiodic Markov Chain because (7) does not converge, this definition guarantees the[2] existence of a unique stationary distribution vector q which becomes the PageRank vector. Given this results a primitive stochastic matrix T which entries are t_{pq} for the algorithm is created:

$$t_{pq} = \begin{cases} \frac{1}{|O_q|} & is\ there\ is\ a\ link\ \ from\ q\ to\ \ p \\ 0 & otherwise \end{cases} \tag{8}$$

[2] A Markov chain is periodic if the chain can return to the state only at multiples of some integer larger than 1.
< http://www.math.uah.edu/stat/markov/Periodicity.html >

After developing,

with $E = \frac{eet}{n}$ as a perturbation matrix, the matrix computation can be normalized to a stationary vector by calculating the powers of the transition matrix, this stationary vector shows the way pages are organized according to the Page Rank algorithm. [8]

When the transition matrix T has in any of its rows all entries as zero that illustrates there is a page with not forward links that means a dangling node exists. Using a Markov Chain model, it could not be possible for estimation. One way of handling with this problem is to replace all rows with e/n, where e is a row vector of all ones and n is the order of the matrix [10].

3.1. PAGERANK EXAMPLE CALCULATION

Figure 6 shows an example based on a small web which consists of three pages A, B and C; where page C links to the pages A and B, page B links to page C and page A links to page B (3 webpages i.e N=3). In this exercise, we set the damping factor d to 0.7.

Changing the damping factor affects the PageRank result, but it does not have any influence on its fundamental principles.

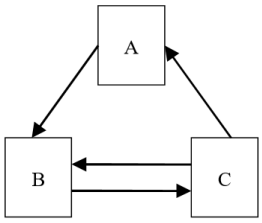

Figure 6 – Network Example

PageRank calculations are expressed in the flowing lines:

$$PR(A) = (1-d)*\left(\frac{1}{N}\right) + d*\left(\frac{PR(C)}{2}\right)$$

$$PR(B) = (1-d)*\left(\frac{1}{N}\right) + d*\left(\frac{PR(A)}{1} + \frac{PR(C)}{2}\right)$$

$$PR(C) = (1-d)*\left(\frac{1}{N}\right) + d*\left(\frac{PR(B)}{2}\right)$$

Previous equations can be simplified as follow:

$$PR(A) = 0.1 + 0.35 * PR(C)$$

$$PR(B) = 0.1 + 0.70 * PR(A) + 0.35 * PR(C)$$

$$PR(C) = 0.1 + 0.70 * PR(B)$$

After solving the above system of linear equations, we get:

$$PR(A) = 0.2314$$
$$PR(B) = 0.3933$$
$$PR(C) = 0.3753$$

These results highlight how the importance of the page B increases due to the reception of different connections from page A and page C.

4. METHODOLOGY

As it was seen in Section 2, PageRank has many approaches. This section shows how this algorithm can identify the referral pattern within the patients in a health company, by using the physician specialties as nodes.

The main goal is to understand the group's referral network and its primary characteristics through the importance of its nodes (i.e medical specialties). For this purpose, PageRank algorithm is going to be used, its objective as it has been show previously, is to search the relative importance of the specialties given the reference that they send and receive. As a result, it is expected to obtain new insights about the patient journey, identify which are the network's points with a higher susceptibility of causing disruptions in the usual route of the patient, and because of this, it will be possible to carry out a more assertive management of the network guarantying its quality and needed capability for offering an excellent service to the patient.

The followed methodology was the use of Markov Chains and the original iterative process of PageRank as it was explained in Section 3

The software used for this exercise was **R**, applying *igraph* [11], *parallel* [12] and *arules* [13] packages.

4.1. DATA SET

The data has information related with patient's events, this means, their medical appointments, medical exams and surgeries made between 2015 and 2016. The analysis was made with 12.321.989 records, which correspond to 1.011.949 patients. The structure of the transactional is the following:

Data Set features:

	ID_Doente	Date	Unidade	Grupo_Rubrica
1	12	05-ene-2016	CCC	Patologia Clinica
2	12	15-ene-2016	HCD	Ginecologia-Obstetrícia
3	12	20-ene-2016	HCIS	Otorrinolaringologia
4	12	30-ene-2016	CCC	Imagiologia

Table 1 - Data Structure

- ID_Doente: Patient Id.
- Date: Medical appointment date.
- Unidade: Medical center location.
- Grupo_Rubrica: Treated medical specialty.

4.2. APRIORI APPROACH

In computer science and data mining, Apriori is a classic algorithm for learning association rules which is a method for discovering interesting relations between variables in large database. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website visit frequency).

As part of the obtained results after applying Apriori algorithm we have the more frequent items (medical specialties): Patologia Clinica, RX Convencional, Urgencia General and Ecografía are the four most popular specialties in the transactional information. See 6.

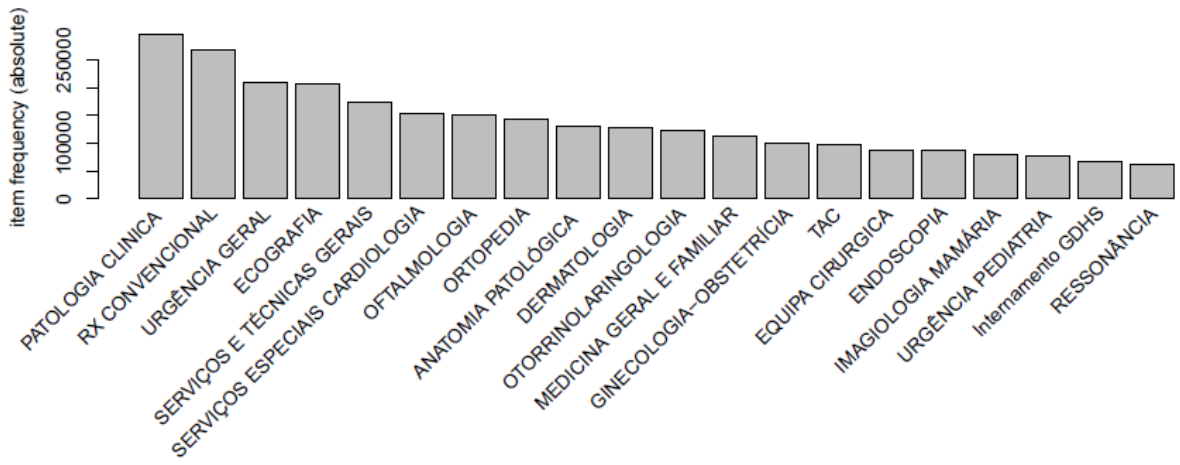

Figure 7 - Frequent Item Set

The main result of this study is the list of the most relevant relations between users of the medical network. In table 2 these relations are sort based on the confidence where the rule {Bloco Operatorio Central}$->${Equipa Cirurgica} is highlighted. This rule, appears in 5% of the total visits. Additionally, in 99% of the cases where Bloco Operatorio Central appears as the antecedent, Equipa Cirurgica is founded as the consequence, presenting this relation as a deterministic one in the data.

In the same way, the remaining rules can be analyzed not as deterministic but with a high degree of certainty in their occurrence. As an example of this group we found the following rule: {Servicios e Tecnicas Gerais, Servicios Especiais Cardiologia} $->$ {Patologia Clinica} which is presented in 5% of the visits and has a degree of certainty of 87% .

For an association rule X $->$ Y, the lift is defined as its performance on predicting cases with a higher probability with respect to the entire population. A lift value equal to 1, means that X and Y are independent. If it is higher than 1, it means that X and Y are positively correlated. Contrary to this, if the lift has a value lower than 1, it means that X and Y are negatively correlated.

As it is shown in Figure 8, rules with a higher lift conforms by themselves a cluster. However, there are some rules to be considered.

The specialty that appears the most in the rules as consequent is PATOLOGIA CLINICA with 12 rules over the top 30 ordered by confidence. For this consequent, the antecedent that increase the most the probability to appear is {SERVIÇOS E TÉCNICAS GERAIS, SERVIÇOS ESPECIAIS CARDIOLOGIA} with a lift of 2.53 and {SERVIÇOS ESPECIAIS CARDIOLOGIA, URGÊNCIA GERAL} with a lift of 2.50. The second specialty that appears the most is ECOGRAFIA with 4 rules, here, the antecedent that increase the most the probability to appear is {GINECOLOGIA-OBSTETRÍCIA, PATOLOGIA CLINICA} with a lift of 3.27. As we can see Patologia clinica is a specialty that appears in both sides of the rules: having a high performance to be predicted and a great performance predicting other specialties.

In the consequent, there are some other remarkable specialties such as GINECOLOGIA-OBSTETRÍCIA that has a lift of 6.01 when {ANATOMIA PATOLÓGICA, ECOGRAFIA} appears as a antecedent. Also ANATOMIA PATOLÓGICA with a lift of 4.88 when {ECOGRAFIA, GINECOLOGIA-OBSTETRÍCIA} is the antecedent. For these 2 rules, the specialties are the same, for this reason, we can conclude bases on the lift value that GINECOLOGIA-OBSTETRÍCIA is more predictable among the two other specialties.

lhs		rhs	support	con dence	lift	count
{BLOCO OPERATORIO CENTRAL}	=>	{EQUIPA CIRURGICA}	0.05	0.99	9.80	46190
{BLOCO OPERATORIO CENTRAL,Internamento GDHS}	=>	{EQUIPA CIRURGICA}	0.05	0.99	9.79	45000
{BLOCO OPERATORIO CENTRAL}	=>	{Internamento GDHS}	0.05	0.97	12.33	45326
{BLOCO OPERATORIO CENTRAL,EQUIPA CIRURGICA}	=>	{Internamento GDHS}	0.05	0.97	12.33	45000
{SERVIÇOS E TÉCNICAS GERAIS,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{PATOLOGIA CLINICA}	0.05	0.87	2.53	43709
{SERVIÇOS ESPECIAIS CARDIOLOGIA,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.05	0.86	2.50	46202
{Internamento GDHS}	=>	{EQUIPA CIRURGICA}	0.07	0.83	8.19	56304
{ECOGRAFIA,SERVIÇOS E TÉCNICAS GERAIS}	=>	{PATOLOGIA CLINICA}	0.06	0.82	2.38	47971
{ECOGRAFIA,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.06	0.81	2.35	54055
{EQUIPA CIRURGICA,Internamento GDHS}	=>	{BLOCO OPERATORIO CENTRAL}	0.05	0.80	14.75	45000
{GINECOLOGIA-OBSTETRÍCIA,PATOLOGIA CLINICA}	=>	{ECOGRAFIA}	0.05	0.78	3.27	44804
{RX CONVENCIONAL,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{PATOLOGIA CLINICA}	0.08	0.78	2.27	69037
{ECOGRAFIA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{PATOLOGIA CLINICA}	0.06	0.77	2.23	52333
{RX CONVENCIONAL,SERVIÇOS E TÉCNICAS GERAIS,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.06	0.77	2.23	49524
{ECOGRAFIA,GINECOLOGIA-OBSTETRÍCIA}	=>	{ANATOMIA PATOLÓGICA}	0.06	0.75	4.88	51210
{PATOLOGIA CLINICA,RX CONVENCIONAL,SERVIÇOS E TÉCNICAS GERAIS}	=>	{URGÊNCIA GERAL}	0.06	0.75	3.05	49524
{RX CONVENCIONAL,SERVIÇOS E TÉCNICAS GERAIS}	=>	{PATOLOGIA CLINICA}	0.08	0.74	2.16	66218
{ANATOMIA PATOLÓGICA,GINECOLOGIA-OBSTETRÍCIA}	=>	{ECOGRAFIA}	0.06	0.74	3.07	51210
{RX CONVENCIONAL,SERVIÇOS E TÉCNICAS GERAIS}	=>	{URGÊNCIA GERAL}	0.08	0.73	2.96	64635
{SERVIÇOS E TÉCNICAS GERAIS,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.09	0.71	2.06	75425
{ECOGRAFIA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{RX CONVENCIONAL}	0.06	0.70	2.23	47755
{GINECOLOGIA-OBSTETRÍCIA}	=>	{ANATOMIA PATOLÓGICA}	0.08	0.70	4.55	69459
{ECOGRAFIA,RX CONVENCIONAL}	=>	{PATOLOGIA CLINICA}	0.08	0.70	2.03	66362
{ANATOMIA PATOLÓGICA,PATOLOGIA CLINICA}	=>	{ECOGRAFIA}	0.06	0.70	2.91	51318
{ANATOMIA PATOLÓGICA,ECOGRAFIA}	=>	{PATOLOGIA CLINICA}	0.06	0.70	2.03	51318
{ANATOMIA PATOLÓGICA,ECOGRAFIA}	=>	{GINECOLOGIA-OBSTETRÍCIA}	0.06	0.70	6.01	51210
{PATOLOGIA CLINICA,SERVIÇOS E TÉCNICAS GERAIS}	=>	{URGÊNCIA GERAL}	0.09	0.70	2.84	75425
{Internamento GDHS}	=>	{PATOLOGIA CLINICA}	0.05	0.70	2.02	47194
{PATOLOGIA CLINICA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{RX CONVENCIONAL}	0.08	0.70	2.22	69037
{IMAGIOLOGIA MAMÁRIA}	=>	{ECOGRAFIA}	0.06	0.69	2.88	55212

Table 2 - Association Rules Top30

To find patterns in an easier way, a graph of the relations is made (See figure 8). Each circle corresponds to the top 12 of the relationships found using the Apriori algorithm ordered by the confidence. The elements of the rules are presented as text and the orientation of the arrows indicates the order of the rule. Red color indicates a greater lift while a larger size indicates greater support.

As a first conclusion, two main groups are identified. The first one involves: Urgencia Geral, Serviços Especiais Cardiologia, Palogia Clinica, Serviços e Tecnicas Gerais, Ecografia, Rx Convercional and Ginecologia Obstetricia. On the other hand, the second group has specialties as: Bloco Operatorio Central, Internamento GDHS and Equipa Cirurguca which have a higher frequency as a group when are compared with different specialties. When analysis of table 2 is add, it is founded that the second main group is related with the first one through the relation {Internamento GDHS} $->$ {PATOLOGIA CLINICA} with a support of 5% and confidence of 70%.

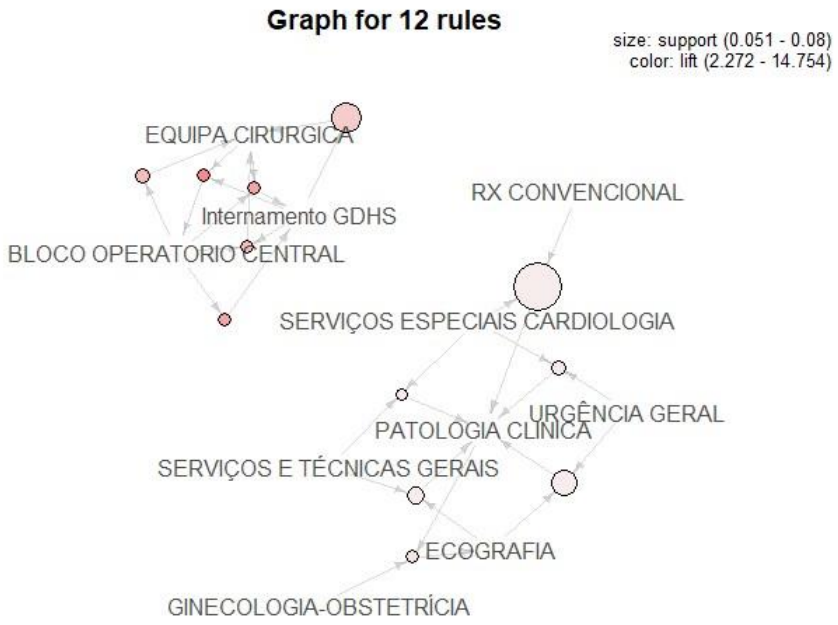

Figure 8 - Rules Graph Top 12

4.3. PAGERANK-GENERAL RESULTS

The way nodes and links were obtained based on the transactional data is explained in figure 9. Nodes are related with Grupo_Rubrica elements and links are created per each user if the patient's visit is presented on 30 days or less. For the example shown in 9 the detail of a patient (ID_Doente=1) with the information of his 4 visits is explored: Connections between Patologia Clinica $->$ Ginecologia-Obstetricia, Patologia Clinica $->$ Otorriniolaringologia and Ginecologia-Obstetricia $->$ Otorrinolaringologia are created because they happened in a time interval lower than 30 days. It is remarkable that all relations have the same weight no matter if the visits are not developed on continuous periods.

ID_Doente	Date	Grupo_Rubrica
1	05-ene-2016	Patologia Clinica
1	15-ene-2016	Ginecologia-Obstetrícia
1	20-ene-2016	Otorrinolaringologia
1	30-Feb-2016	Imagiologia

Figure 9 - Data Base PageRank

According to the methodology explained above, our final dataset has 16.722.470 connections founded between nodes. PageRank algorithm is applied to this final network-shaped dataset, using markov chain approach as it was seen in section 3.

On table 3 is presented the top 15 of the most important nodes according to Pagerank in the complete network built. In the same way tables 4 and 5 show the relations with their adjacent edges, this means, the relations that get out from each node and get in it respectively.

The most important node is PATOLOGIA CLINICA, which is highlighted because of its PageRank score, its Out-degree and In-degree, the highest in all cases. The second most important node is SERVIÇOS E TECNICAS GERAIS, which differs from the first one in its Out-degree. This means that patients are sent to this node by other nodes of high importance.

Node	PageRank
PATOLOGIA CLINICA	7.30
SERVIÇOS E TÉCNICAS GERAIS	5.00
ORTOPEDIA	4.44
ECOGRAFIA	4.35
ANATOMIA PATOLÓGICA	3.87
RX CONVENCIONAL	3.75
GINECOLOGIA-OBSTETRÍCIA	2.98
OTORRINOLARINGOLOGIA	2.96
SERVIÇOS ESPECIAIS CARDIOLOGIA	2.75
MEDICINA GERAL E FAMILIAR	2.67
TRATAMENTO MEDICINA FISICA E REABILITAÇÃO	2.12
TAC	1.95
OFTALMOLOGIA	1.91
DERMATOLOGIA	1.88
URGÊNCIA GERAL	1.75

Table 3 - Top 15 Nodes-Page Rank

Node	OUT-Degree
1 PATOLOGIA CLINICA	141806
2 RX CONVENCIONAL	84463
3 ECOGRAFIA	75971
4 GINECOLOGIA-OBSTETRÍCIA	61927
5 MEDICINA GERAL E FAMILIAR	57812
6 URGÊNCIA GERAL	56854
7 ORTOPEDIA	56640
8 SERVIÇOS ESPECIAIS CARDIOLOGIA	54366
9 ANATOMIA PATOLÓGICA	44086
10 SERVIÇOS E TÉCNICAS GERAIS	43729

Table 4 - Top 10 Out-Degree

Node	IN-Degree
1 PATOLOGIA CLINICA	118474
2 SERVIÇOS E TÉCNICAS GERAIS	76517
3 ECOGRAFIA	74842
4 ORTOPEDIA	73130
5 ANATOMIA PATOLÓGICA	67623
6 RX CONVENCIONAL	60331
7 MEDICINA GERAL E FAMILIAR	48006
8 GINECOLOGIA-OBSTETRÍCIA	46549
9 SERVIÇOS ESPECIAIS CARDIOLOGIA	45012
10 OTORRINOLARINGOLOGIA	44799

Table 5 - Top 10 In-Degree

In the following table (6), relations with the higher frequency that were found in the network are shown. Following this idea, RX CONVENCIONAL -> ORTOPEDIA is the relation with the superior value. The complete path which a higher frequency according to the weights of the relations is RX CONVENCIONAL<->ORTOPEDIA->PATOLOGIA CLINICA<->MEDICINA GERAL E FAMILIAR->ECOGRAFIA<->GINECOLOGIA-OBSTETRCIA->ANATOMIA PATOLÓGICA.

Grupo_Rubrica_From	Grupo_Rubrica_To	Weight
RX CONVENCIONAL	ORTOPEDIA	15327
MEDICINA FÍSICA E REABILITAÇÃO	TRATAMENTO MEDICINA FISICA E REABILITAÇÃO	14063
PATOLOGIA CLINICA	MEDICINA GERAL E FAMILIAR	11874
PATOLOGIA CLINICA	ECOGRAFIA	11620
MEDICINA GERAL E FAMILIAR	PATOLOGIA CLINICA	11544
GINECOLOGIA-OBSTETRÍCIA	PATOLOGIA CLINICA	9766
ECOGRAFIA	GINECOLOGIA-OBSTETRICIA	9476
ECOGRAFIA	PATOLOGIA CLINICA	9172
PATOLOGIA CLINICA	GINECOLOGIA-OBSTETRÍCIA	9072
GINECOLOGIA-OBSTETRÍCIA	ANATOMIA PATOLÓGICA	8529
GINECOLOGIA-OBSTETRÍCIA	ECOGRAFIA	8496
ORTOPEDIA	RX CONVENCIONAL	8328
TRATAMENTO MEDICINA FISICA E REABILITAÇÃO	MEDICINA FÍSICA E REABILITAÇÃO	8112
RX CONVENCIONAL	PATOLOGIA CLINICA	7175
EQUIPA CIRURGICA	SERVIÇOS E TÉCNICAS GERAIS	7109
URGÊNCIA GERAL	PATOLOGIA CLINICA	6860
PATOLOGIA CLINICA	SERVIÇOS ESPECIAIS CARDIOLOGIA	6802
OTORRINOLARINGOLOGIA	SERVIÇOS ESPECIAIS OTORRINOLARINGOLOGIA	6778
SERVIÇOS E TÉCNICAS GERAIS	ORTOPEDIA	6708
ANATOMIA PATOLÓGICA	PATOLOGIA CLINICA	6260

Table 6 - Top 20 Relations

Figure 10 presents the network's structure of the formed relation's and nodes with lower Out-Degree are pointed up. This is, nodes where the patients end their medical services route end: CIRURGIA CARDIACA, MEDICINA E CIRURGIA ESTÉTICA, SERVIÇOS ESPECIAIS, HEMATO−ONCOLOGIA, REABILITAÇÃO COGNITIVA, MEDICINA ANTI−ENVELHECIMENTO y URGÊNCIA REUMATOLOGIA are highlighted.

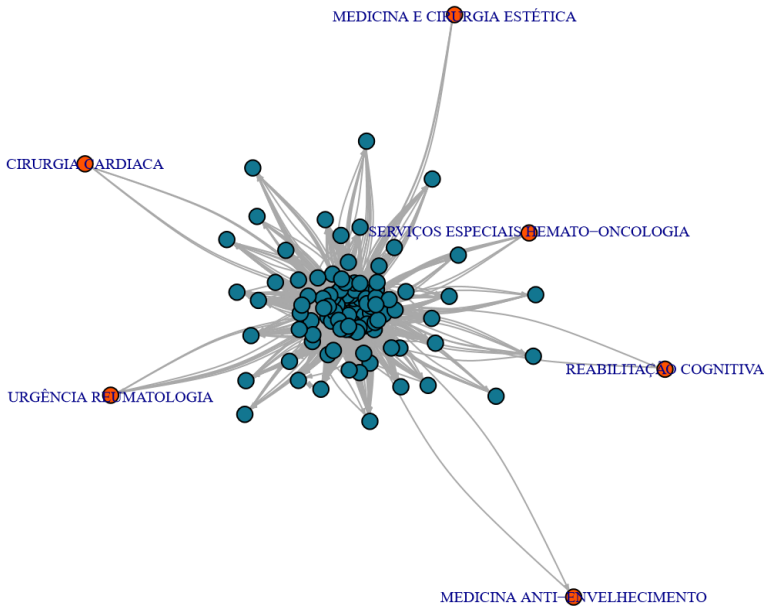

Figure 10 – Dead Nodes

To obtain a better understanding of the main elements, the network is filtered based on the top 15 nodes established in the general network (Table 3) and supposing that the weight of each relation is higher than the average.

The following figure (11) shows the resulting network, where the relationships between the most relevant elements can be observed in a more clearly way: orange nodes which are three in total represent the most important nodes in the new network. See table 7.

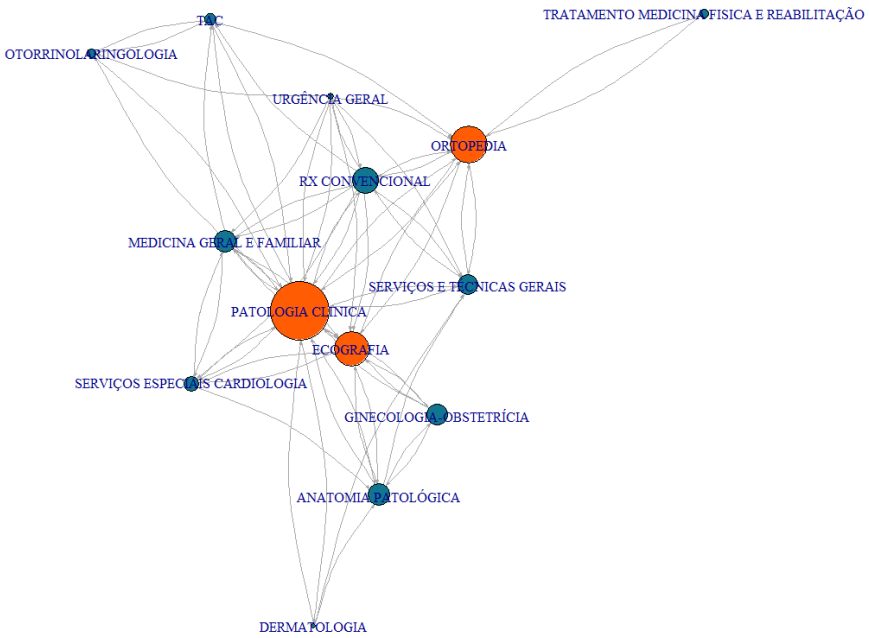

Figure 11 – Most Important Network

OFTALMOLOGIA node is discarded from the analysis since the absence of the previous conditions needed. It can be observed that the actual distribution changes with respect to the general one. Particularly, SERVIÇOS E TÉCNICAS GERAIS node changes its position from 2nd to 8th in the new network, this, due to the connections with less important nodes it had.

Node	PageRank
PATOLOGIA CLINICA	19.56
ORTOPEDIA	12.51
ECOGRAFIA	11.65
RX CONVENCIONAL	8.66
MEDICINA GERAL E FAMILIAR	7.34
ANATOMIA PATOLÓGICA	7.20
GINECOLOGIA-OBSTETRÍCIA	7.06
SERVIÇOS E TÉCNICAS GERAIS	6.81
SERVIÇOS ESPECIAIS CARDIOLOGIA	5.12
TAC	3.94
TRATAMENTO MEDICINA FISICA E REABILITAÇÃO	3.16
OTORRINOLARINGOLOGIA	3.02
URGÊNCIA GERAL	2.28
DERMATOLOGIA	1.70

Table 7 – PageRank Most Important Network

For the analysis of the new network, a Community detection based on edge betweenness methodology is applied, It involves iterative removal of edges from the network to split it into communities, the edges removed being identified using one of a number of possible "betweenness" measures, and second, these measures are, crucially, recalculated after each removal [14]. The resulting dendrogram (Figure 12) is presented next describing the cluster's creation process.

The first nodes that establish a cluster are DERMATOLOGIA - ANATOMIA PATOLÓGICA, TAG - OTORRINOLARINGOLOGIA and SERVIÇOS ESPECIAIS CARDIOLOGIA – MEDICINA GERAL E FAMILIAR. Based on these results, health company can identify critical points for patient attention, in other words, knowing that a person visits one of the specialties of a cluster it should be prepare for client's request for medical consultation of the other specialty since it is very likely that the patient requests it.

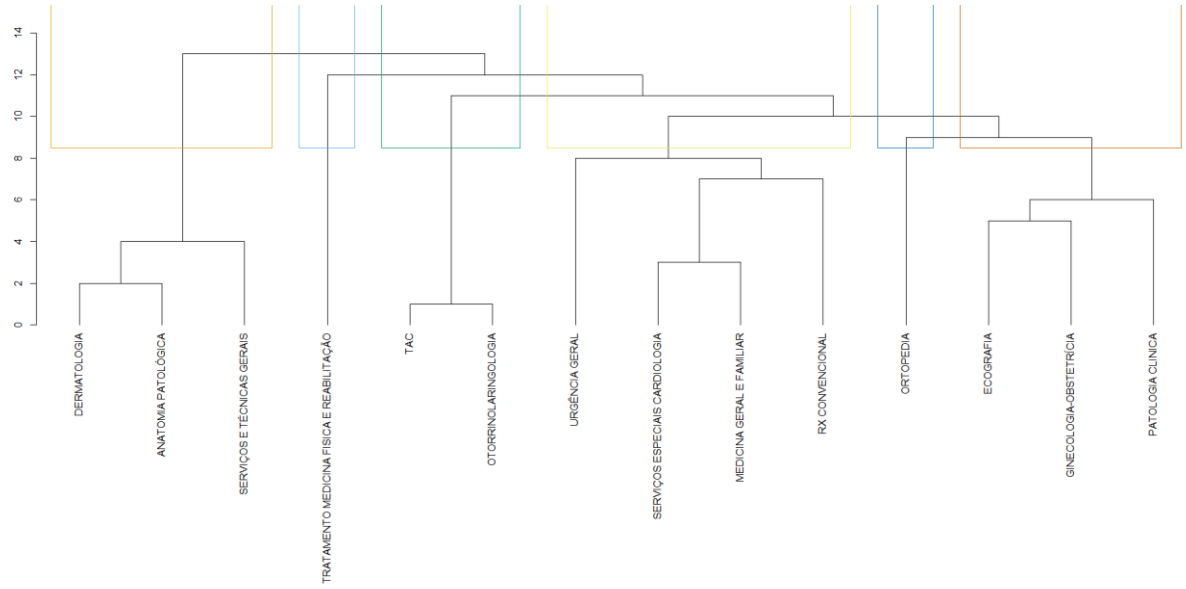

Figure 12 – Network´s Dendrogram

Figure 13 shows the final result with the clusters identified. Here, the outline, point out the individulas that belong to each group, black arrows represent the connections whitin clusters while the red ones the links between them.

As a result, six clusters are observed: two of them correspond to groups created with only one individual (TRATAMENTO MEDICINA FISICA Y REABILITAÇÃO and ORTOPEDIA).

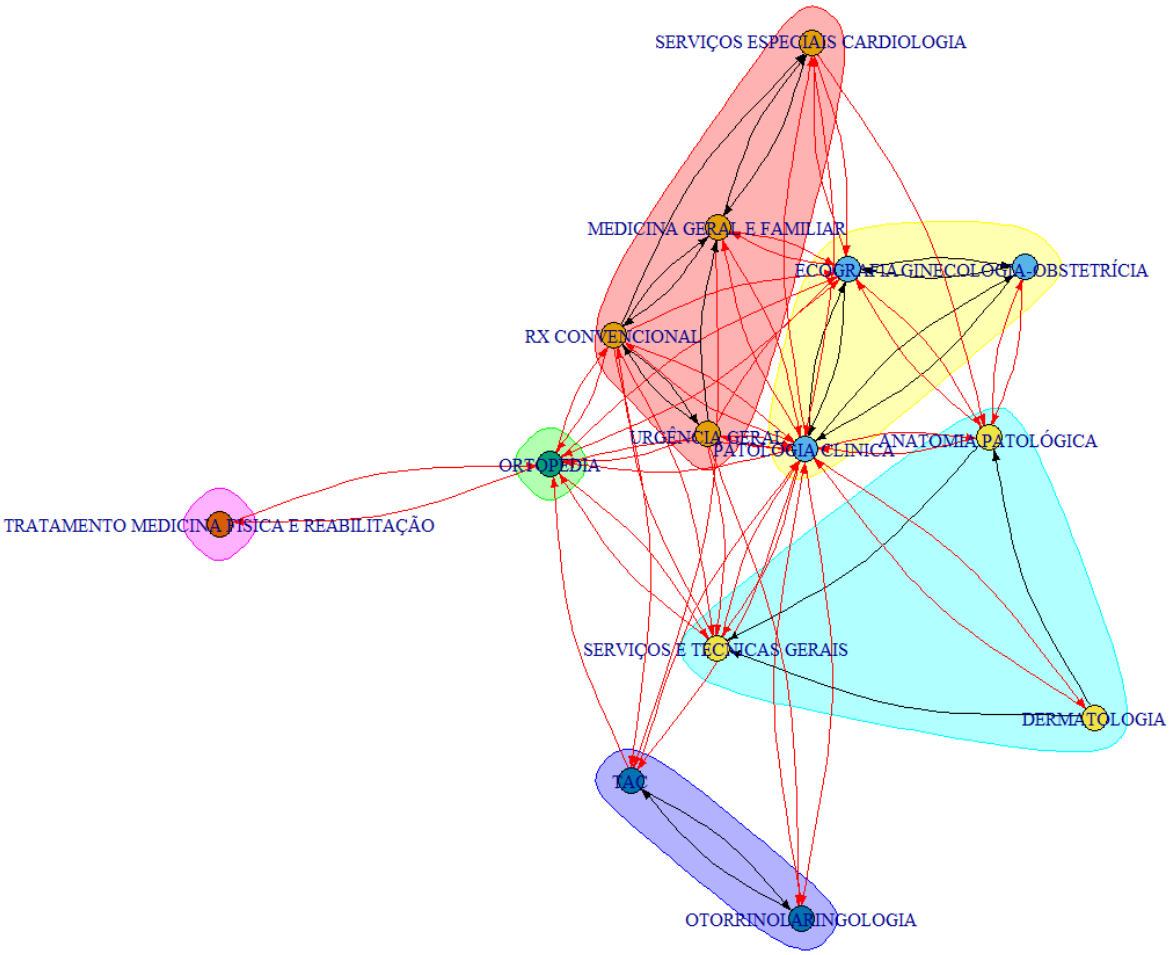

Figure 13 – Network Clustering

After analyzing the resulting big numbers, the network is splitted based on the weight of each link, this is, the number of times that the relation was found. By doing this, a complete vision of the network's elements is achieve, even for the relations which frequency is not high. In the following section 4.4 results are detailed.

4.4. PAGERANK-DEAD NODES BY QUARTILES OF NUMBER OF OCCURRENCES

As it was presented in Table 6, the most frequent relation found in the network was RX CONVENCIONAL -> ORTOPEDIA, with 15.237 appearances (link's weight). This value is the highest in the weight distribution. Based on the frequency of relationships that were found a split is made according to the quartiles.

On table 8, the distribution on the total number of links found in the relations is presented. Based on it, some comments are displayed:

- 25% of the relations were found between one or two times. In average, each relation appears 10 times.
- It is evident a skewness distribution for the frequencies: the median has a value of 10 while the maximum value reaches 15.237. Top 20 relation's frequency is shown in table 6, previously mentioned.
- It is remarkable that even if a relation's cluster does not appear with a numerous frequency, each network's element (node) could be analyzed and categorized according to the internal behavior of the group. This means, each node allows to be sorted based on its relative importance in the cluster. PageRank is used to discover these elements and the nodes where the network's flow stops, the analysis must be made for each group obtained by the quartile-based division. Results for each quartile are presented next.

Quantile Weight				
0%	25%	50%	75%	100%
1	2	10	58	15327

Table 8 - Quantile Distribution

4.4.1. First Quartile ($Weigth <= 2$)

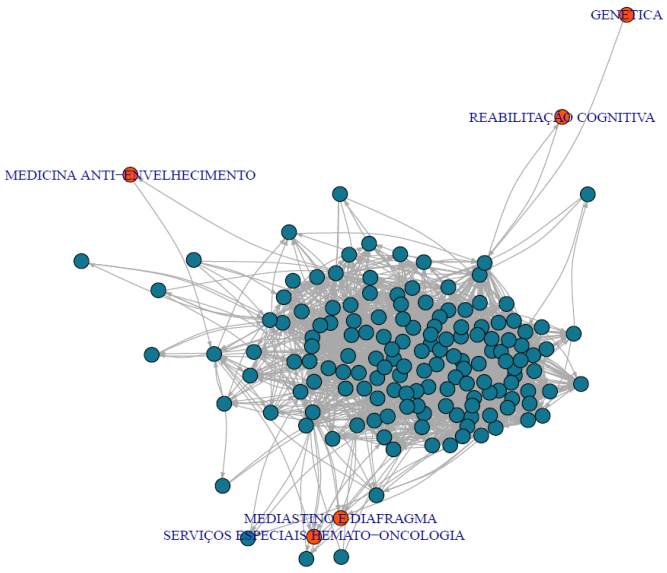

Figure 14 – Dead Nodes First Quartile

Node	PageRank
DIÁRIAS	1.33
MEDICINA NUCLEAR	1.26
ONCOLOGIA	1.23
URGÊNCIA OTORRINOLARINGOLOGIA	1.18
CHECK-UP	1.17
CIRURGIA MAXILO-FACIAL	1.15
DIAGNÓSTICO E TERAPÊUTICA VASCULAR	1.09
URGÊNCIA ORTOPEDIA	1.05
MEDICINA INTERNA	1.04
TEMP	1.03

Table 9 - Top 10 Nodes First Quartile-Page Rank

4.4.2. Second Quartile ($2 < Weigth <= 10$)

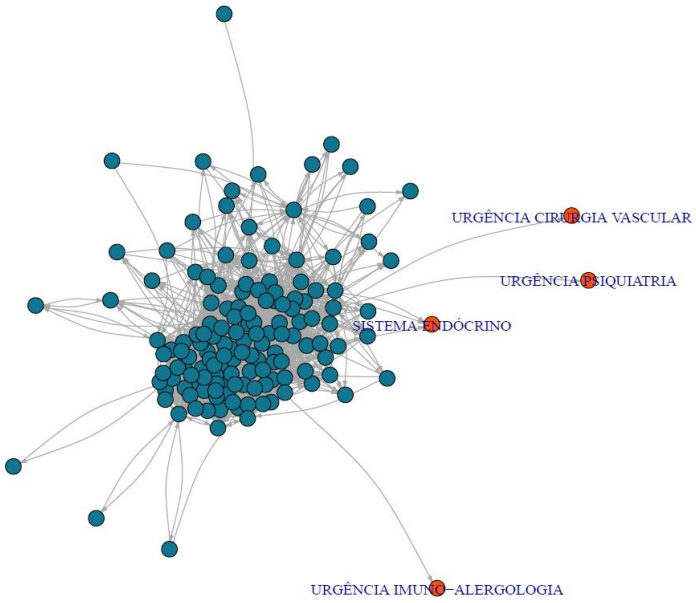

Figure 15 – Dead-Nodes Second Quartile

Node	PageRank
URGÊNCIA ORTOPEDIA	1.47
ANATOMIA PATOLÓGICA	1.45
Internamento GDHS	1.26
EQUIPA CIRURGICA	1.25
BLOCO OPERATORIO CENTRAL	1.20
PSICOLOGIA	1.17
CIR. PLÁSTICA E RECONSTRUTIVA	1.16
CIRURGIA GERAL	1.15
NEUROLOGIA	1.14
SERVIÇOS ESPECIAIS UROLOGIA	1.13

Table 10 - Top 10 Nodes Second Quartile-Page Rank

4.4.3. Third Quartile ($10 < Weigth <= 58$)

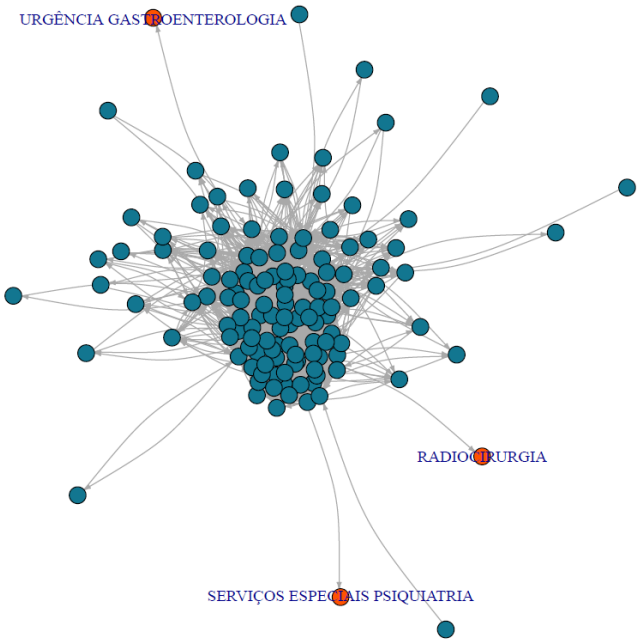

Figure 16 - Dead-Nodes Third Quartile

Node	PageRank
PATOLOGIA CLINICA	2.69
ANATOMIA PATOLÓGICA	1.99
RX CONVENCIONAL	1.58
SERVIÇOS ESPECIAIS CARDIOLOGIA	1.55
UROLOGIA	1.51
SERVIÇOS E TÉCNICAS GERAIS	1.51
URGÊNCIA GERAL	1.51
MEDICINA FÍSICA E REABILITAÇÃO	1.46
ECOGRAFIA	1.46
GINECOLOGIA-OBSTETRÍCIA	1.45

Table 11 - Top 10 Nodes Third Quartile-Page Rank

4.4.4. Fourth Quartile($58 < Weigth <= 15327$)

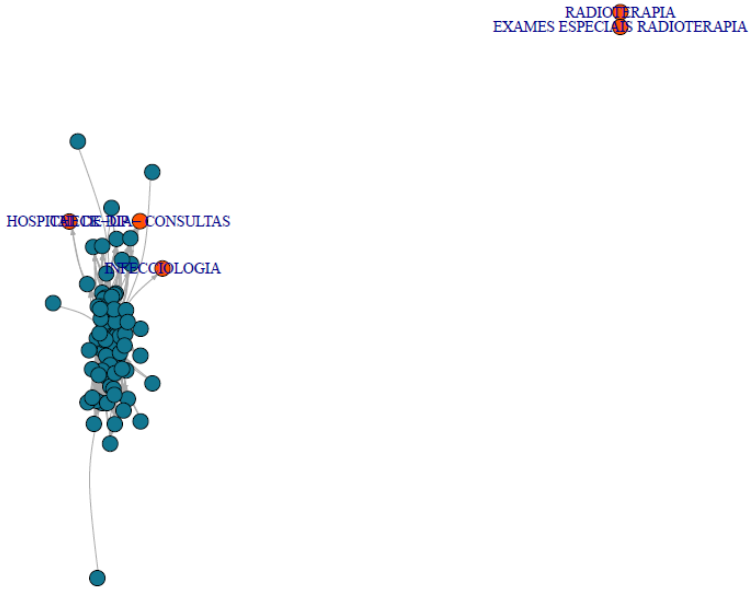

Figure 17 - Dead-Nodes Fourth Quartile

Node	PageRank
PATOLOGIA CLINICA	7.75
SERVIÇOS E TÉCNICAS GERAIS	5.59
ORTOPEDIA	4.86
ECOGRAFIA	4.58
ANATOMIA PATOLÓGICA	3.99
RX CONVENCIONAL	3.87
OTORRINOLARINGOLOGIA	3.12
GINECOLOGIA-OBSTETRÍCIA	3.08
MEDICINA GERAL E FAMILIAR	2.76
SERVIÇOS ESPECIAIS CARDIOLOGIA	2.68

Table 12 - Top 10 Nodes Fourth Quartile-Page Rank

Detailing the network's structure differentiating it by the frequency of connections allows to design and establish strategies for each element that intervenes in the path followed by the patient through the studied medical network. In addition to the general study of the connections, on section 4.5 detail of patient's characteristics are shown, particularly their gender and location.

4.5. PAGERANK BY GENDER AND LOCATION

As it is observed on figure 18, nodes with more influence on women than men are Patologia Clinica,Ecografia y Anatomia Patológica. For men, nodes with higher importance are: Serviços e Tecnicas Gerais, Ortopedia, Rx Convencional, Otorrinolaringologia, Serviços Especiais Cardiologia y Medicina Geral.

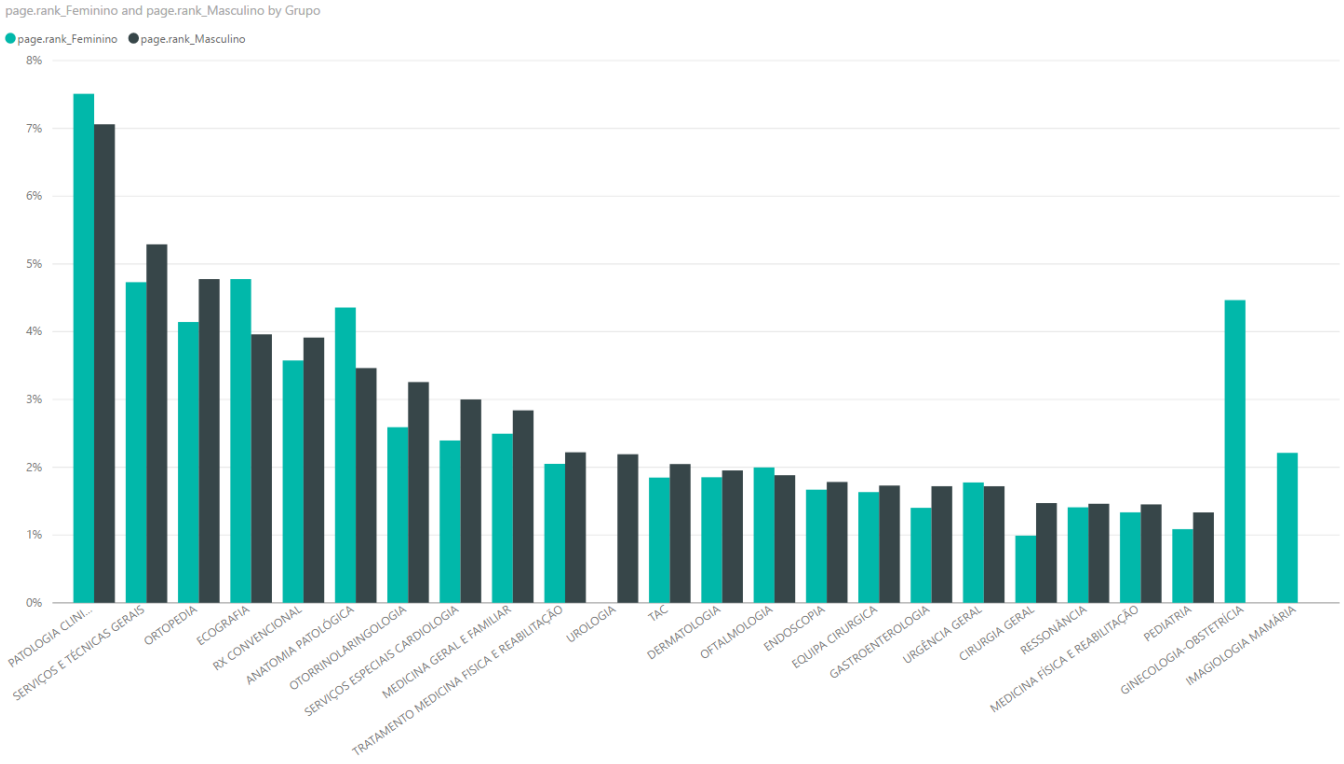

Figure 18 - Importance by Gender

On figure 19, it is recognized the relevance of the nodes according to different geographic areas, districts as Santarem, Aveiro and Porto stand out having a different distribution in the arrangement of the nodes. For Santarem, the most important node is Ortopedia while Aveiro and Porto present Serviços e Tecnicas Gerais as the most significant.

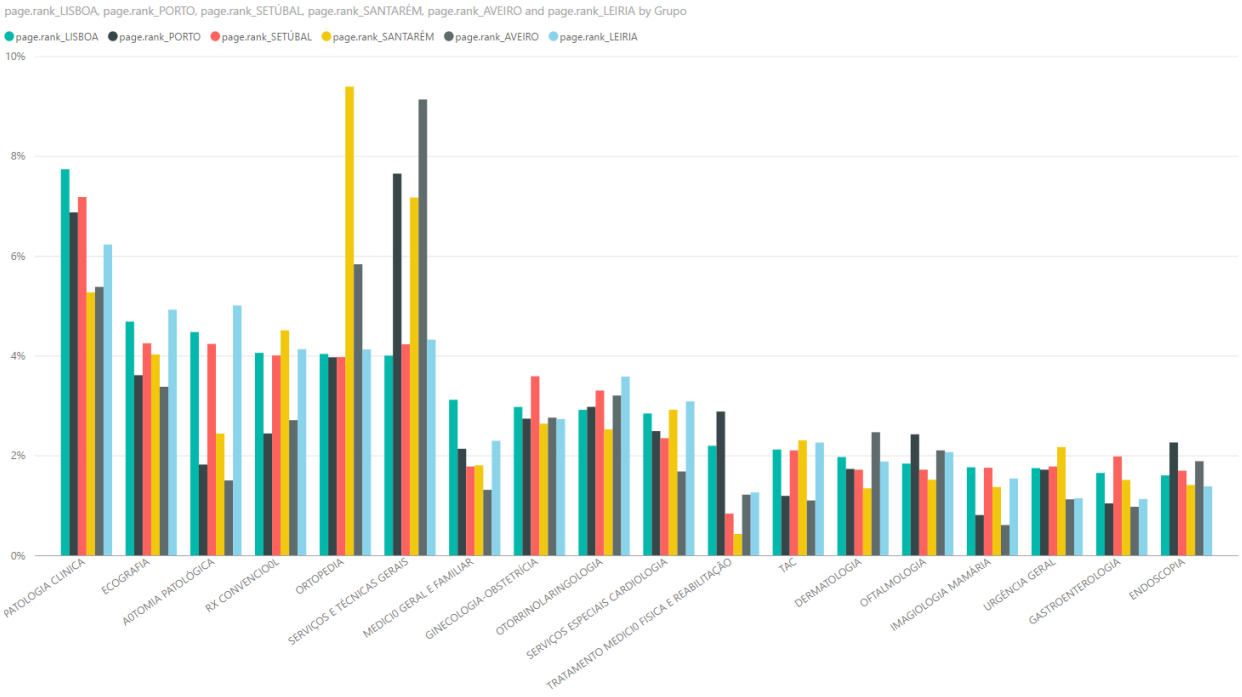

Figure 19 - Importance by Location

5. CONCLUSIONS

Using a network as the basis for studying a transactional database allows to identify main patterns based on the existing connections between elements and visualize them in a more tangible way. Working with tools as the PageRank algorithm is useful to identify the most important elements of the group of elements in the analysis. In medicine, this methodology was applied on medical specialties, treating them as the nodes of a network where connections between them represent the visit of a specific patient.

Applying this methodology and identifying elements such as dead nodes guarantee the application of clustering creating distances based on the connections and PageRank algorithm as a sorting tool. The last one was useful thanks to its flexibility and ease of calculation. Results are presented by the distributions of the relation's weights, patients' descriptive variables (genre) genre and companies' information (Location) allowing a deeper understanding of the patient's journey

The Results of this thesis give descriptive analysis considering the physician referral network found in the transactional data, this means that It is useful for generate strategies accompanied by a high knowledge of the characteristics of the health company. In terms of computing costs, convert the transactional data into a network structure involve a big challenge, the solution was found in the use of "parallel computing" decreasing processing time in a considerable way by increasing the machine resources in a vertical mode.

After understanding an applying PageRank algorithm, it is remarkable how useful is for obtaining weighted and less subjective rankings. Its applicability allows it to be a differential tool in terms of quality and accuracy solutions not only listing webpages in terms of their importance for search engines but in any problem which have a link structure and can be represented as a network.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

PageRank algorithm is based on the idea, in which a random surfer is imagined navigating through the web as a random walker, he follows the edges in an uniformly way with a given probability α, but randomly he will jump to any page with probability $1-\alpha$. The uniformly distribution do not allow derive PageRank values tailored to particular interests. For example, the company would be interested in some particular specialty due its financial status or how new it is, and wants this specialty be ranked higher than other ones. For resolving this new goal, it is necessary to apply an algorithm called Topic-specific PageRank, where the teleporting to a random web page is chosen *non-uniformly.* We can express the PageRank as a solution for:

$$\overline{T} = \alpha T + (1-\alpha)E$$

$T = e/n$, where e is a row vector of all ones and n is the order of the matrix

Topic-specific PageRank can be computed by increasing the bias in the calculation using a nonuniform matrix *T.* The results of doing this is that the effect of certain categories can be amplified.

7. BIBLIOGRAPHY

[1] Page,Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry, *The PageRank Citation Ranking:Bringing Order to the Web*, ＜http://google.stanford.edu:/∶backrub/pageranksub.ps＞.

[2] Kumar,Ravi and Sivakumar, D , *Algortihms and Models for the Web Graph*, 7^{th} International Workshop, WAW 2010, Stanford, CA, USA. Springer.2010

[3] Page,Lawrence and Brin, Sergey, *Reprint of: The anatomy of a large-scale hypertextual web search engine*, ScienceDirect, Computer Networks, 2012.

[4] Wang, Jing and Liu, Zhijing and Zhao, Hui, *Group Recommendation Based on the PageRank*, Journal of Networks, Vol 7, 2012.

[5] Newton, Paul and Mason, Jeremy and Behtel, Kelly and Bazhenova, Lyudmilla and Nieva, Jorgeand Kuhn, Peter, *A stochastic Markov Chain Model to Describe Lung Cancer Growth and Metastasis*, PLosOne, Vol 7, 2012.

[6] Lazova, Verica and Barskanov,Lasko , *PageRank Approach to Ranking National Football Teams*, Cornell University , 2015.

[7] Allesina, Stefano and Pascual, Mercedes , *Googling Food Webs: Can an Eigenvector Measure Species' Importance for Coextinctions?* ScienceDaily. ScienceDaily,2009.

[8] Kumar, Ravi and Kwang Leng, Alex G and Kwang Leng, Alex G , *Application of Markov Chain in the PageRank Algorithm*

[9] Langville, A.N and Meyer, C.D , *The use of the Linear Algebra by Web Search Engines* Bulletin of the International Linear Algebra, No. 23,2004.

[10] Langville, A.N and Meyer, C.D , *Deeper Inside PageRank* Internet Mathematics, Vol 1, Issue 3,pp 335-380. 2004.

[11] Csardi, Gabor and Nepusz,Tamas , *The igraph software package for complex network research* InterJournal, Complex Systems,pp 1695, 2006. ＜http://igraph.org＞

[12] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

[13] Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik (2017). arules: Mining Association Rules and Frequent Itemsets. R package version 1.5-3. https://CRAN.R-project.org/package=arules.

[14] M Newman and M Girvan: Finding and evaluating community structure in networks, Physical Review E 69, 026113 (2004)

8. ANNEXES

8.1. ASSOCIATION RULES TABLE

	lhs		rhs	support	confidence	lift	count
1	{BLOCO OPERATORIO CENTRAL}	=>	{EQUIPA CIRURGICA}	0.054	0.993	9.796	46190
2	{BLOCO OPERATORIO CENTRAL,Internamento GDHS}	=>	{EQUIPA CIRURGICA}	0.052	0.993	9.795	45000
3	{BLOCO OPERATORIO CENTRAL}	=>	{Internamento GDHS}	0.053	0.974	12.335	45326
4	{BLOCO OPERATORIO CENTRAL,EQUIPA CIRURGICA}	=>	{Internamento GDHS}	0.052	0.974	12.333	45000
5	{SERVIÇOS E TÉCNICAS GERAIS,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{PATOLOGIA CLINICA}	0.051	0.872	2.535	43709
6	{SERVIÇOS ESPECIAIS CARDIOLOGIA,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.054	0.861	2.504	46202
7	{Internamento GDHS}	=>	{EQUIPA CIRURGICA}	0.066	0.830	8.189	56304
8	{ECOGRAFIA,SERVIÇOS E TÉCNICAS GERAIS}	=>	{PATOLOGIA CLINICA}	0.056	0.818	2.379	47971
9	{ECOGRAFIA,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.063	0.808	2.348	54055
10	{EQUIPA CIRURGICA,Internamento GDHS}	=>	{BLOCO OPERATORIO CENTRAL}	0.052	0.799	14.754	45000
11	{GINECOLOGIA-OBSTETRÍCIA,PATOLOGIA CLINICA}	=>	{ECOGRAFIA}	0.052	0.785	3.269	44804
12	{RX CONVENCIONAL,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{PATOLOGIA CLINICA}	0.080	0.781	2.272	69037
13	{ECOGRAFIA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{PATOLOGIA CLINICA}	0.061	0.766	2.228	52333
14	{RX CONVENCIONAL,SERVIÇOS E TÉCNICAS GERAIS,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.058	0.766	2.227	49524
15	{ECOGRAFIA,GINECOLOGIA-OBSTETRÍCIA}	=>	{ANATOMIA PATOLÓGICA}	0.060	0.749	4.879	51210
16	{PATOLOGIA CLINICA,RX CONVENCIONAL,SERVIÇOS E TÉCNICAS GERAIS}	=>	{URGÊNCIA GERAL}	0.058	0.748	3.047	49524
17	{RX CONVENCIONAL,SERVIÇOS E TÉCNICAS GERAIS}	=>	{PATOLOGIA CLINICA}	0.077	0.744	2.162	66218
18	{ANATOMIA PATOLÓGICA,GINECOLOGIA-OBSTETRÍCIA}	=>	{ECOGRAFIA}	0.060	0.737	3.072	51210
19	{RX CONVENCIONAL,SERVIÇOS E TÉCNICAS GERAIS}	=>	{URGÊNCIA GERAL}	0.075	0.726	2.958	64635
20	{SERVIÇOS E TÉCNICAS GERAIS,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.088	0.708	2.059	75425
21	{ECOGRAFIA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{RX CONVENCIONAL}	0.056	0.699	2.235	47755
22	{GINECOLOGIA-OBSTETRÍCIA}	=>	{ANATOMIA PATOLÓGICA}	0.081	0.699	4.553	69459
23	{ECOGRAFIA,RX CONVENCIONAL}	=>	{PATOLOGIA CLINICA}	0.077	0.699	2.031	66362
24	{ANATOMIA PATOLÓGICA,PATOLOGIA CLINICA}	=>	{ECOGRAFIA}	0.060	0.699	2.911	51318
25	{ANATOMIA PATOLÓGICA,ECOGRAFIA}	=>	{PATOLOGIA CLINICA}	0.060	0.698	2.028	51318
26	{ANATOMIA PATOLÓGICA,ECOGRAFIA}	=>	{GINECOLOGIA-OBSTETRÍCIA}	0.060	0.696	6.014	51210
27	{PATOLOGIA CLINICA,SERVIÇOS E TÉCNICAS GERAIS}	=>	{URGÊNCIA GERAL}	0.088	0.696	2.836	75425
28	{Internamento GDHS}	=>	{PATOLOGIA CLINICA}	0.055	0.696	2.023	47194
29	{PATOLOGIA CLINICA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{RX CONVENCIONAL}	0.080	0.695	2.223	69037
30	{IMAGIOLOGIA MAMÁRIA}	=>	{ECOGRAFIA}	0.064	0.691	2.878	55212
31	{GINECOLOGIA-OBSTETRÍCIA}	=>	{ECOGRAFIA}	0.080	0.688	2.866	68386
32	{Internamento GDHS}	=>	{BLOCO OPERATORIO CENTRAL}	0.053	0.668	12.335	45326
33	{ECOGRAFIA,URGÊNCIA GERAL}	=>	{RX CONVENCIONAL}	0.052	0.665	2.127	44526
34	{PATOLOGIA CLINICA,SERVIÇOS E TÉCNICAS GERAIS,URGÊNCIA GERAL}	=>	{RX CONVENCIONAL}	0.058	0.657	2.098	49524
35	{PATOLOGIA CLINICA,RX CONVENCIONAL,URGÊNCIA GERAL}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.058	0.655	3.231	49524
36	{ECOGRAFIA,GINECOLOGIA-OBSTETRÍCIA}	=>	{PATOLOGIA CLINICA}	0.052	0.655	1.905	44804
37	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{PATOLOGIA CLINICA}	0.116	0.647	1.881	99265
38	{EQUIPA CIRURGICA}	=>	{Internamento GDHS}	0.066	0.647	8.189	56304
39	{ORTOPEDIA}	=>	{RX CONVENCIONAL}	0.108	0.641	2.049	92394

40	{RX CONVENCIONAL,URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.088	0.640	1.861	75584
41	{SERVIÇOS E TÉCNICAS GERAIS}	=>	{PATOLOGIA CLINICA}	0.126	0.622	1.809	108358
42	{ENDOSCOPIA}	=>	{ANATOMIA PATOLÓGICA}	0.063	0.620	4.038	53812
43	{PATOLOGIA CLINICA,URGÊNCIA GERAL}	=>	{RX CONVENCIONAL}	0.088	0.615	1.967	75584
44	{PATOLOGIA CLINICA,URGÊNCIA GERAL}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.088	0.614	3.029	75425
45	{SERVIÇOS E TÉCNICAS GERAIS}	=>	{URGÊNCIA GERAL}	0.124	0.612	2.492	106501
46	{PATOLOGIA CLINICA,SERVIÇOS E TÉCNICAS GERAIS}	=>	{RX CONVENCIONAL}	0.077	0.611	1.953	66218
47	{MEDICINA GERAL E FAMILIAR}	=>	{PATOLOGIA CLINICA}	0.080	0.610	1.774	68853
48	{TAC}	=>	{PATOLOGIA CLINICA}	0.069	0.609	1.770	59169
49	{SERVIÇOS E TÉCNICAS GERAIS,URGÊNCIA GERAL}	=>	{RX CONVENCIONAL}	0.075	0.607	1.940	64635
50	{ECOGRAFIA}	=>	{PATOLOGIA CLINICA}	0.141	0.588	1.708	121109
51	{EQUIPA CIRURGICA}	=>	{PATOLOGIA CLINICA}	0.059	0.585	1.699	50882
52	{URGÊNCIA GERAL}	=>	{PATOLOGIA CLINICA}	0.143	0.583	1.694	122823
53	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{RX CONVENCIONAL}	0.103	0.576	1.840	88340
54	{GINECOLOGIA-OBSTETRÍCIA}	=>	{PATOLOGIA CLINICA}	0.067	0.575	1.670	57110
55	{TAC}	=>	{RX CONVENCIONAL}	0.065	0.572	1.828	55585
56	{URGÊNCIA GERAL}	=>	{RX CONVENCIONAL}	0.137	0.560	1.790	118043
57	{IMAGIOLOGIA MAMÁRIA}	=>	{PATOLOGIA CLINICA}	0.052	0.559	1.626	44709
58	{ANATOMIA PATOLÓGICA}	=>	{ECOGRAFIA}	0.086	0.558	2.326	73563
59	{ANATOMIA PATOLÓGICA}	=>	{PATOLOGIA CLINICA}	0.086	0.557	1.620	73461
60	{ECOGRAFIA,PATOLOGIA CLINICA}	=>	{RX CONVENCIONAL}	0.077	0.548	1.751	66362
61	{RX CONVENCIONAL,URGÊNCIA GERAL}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.075	0.548	2.701	64635
62	{PATOLOGIA CLINICA,RX CONVENCIONAL}	=>	{URGÊNCIA GERAL}	0.088	0.541	2.205	75584
63	{RX CONVENCIONAL,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{ECOGRAFIA}	0.056	0.541	2.252	47755
64	{EQUIPA CIRURGICA}	=>	{BLOCO OPERATORIO CENTRAL}	0.054	0.531	9.796	46190
65	{PATOLOGIA CLINICA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{ECOGRAFIA}	0.061	0.527	2.197	52333
66	{ANATOMIA PATOLÓGICA}	=>	{GINECOLOGIA-OBSTETRÍCIA}	0.081	0.527	4.553	69459
67	{RX CONVENCIONAL}	=>	{PATOLOGIA CLINICA}	0.163	0.520	1.511	139629
68	{SERVIÇOS E TÉCNICAS GERAIS}	=>	{RX CONVENCIONAL}	0.104	0.511	1.634	89020
69	{ENDOSCOPIA}	=>	{PATOLOGIA CLINICA}	0.051	0.508	1.478	44154
70	{EQUIPA CIRURGICA}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.052	0.508	2.507	44249
71	{URGÊNCIA GERAL}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.124	0.505	2.492	106501
72	{ECOGRAFIA,RX CONVENCIONAL}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.056	0.503	2.815	47755
73	{PATOLOGIA CLINICA,RX CONVENCIONAL}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.080	0.494	2.768	69037
74	{TAC}	=>	{URGÊNCIA GERAL}	0.055	0.488	1.987	47391
75	{PATOLOGIA CLINICA,RX CONVENCIONAL}	=>	{ECOGRAFIA}	0.077	0.475	1.980	66362
76	{PATOLOGIA CLINICA,RX CONVENCIONAL}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.077	0.474	2.339	66218
77	{PATOLOGIA CLINICA}	=>	{RX CONVENCIONAL}	0.163	0.473	1.511	139629
78	{ECOGRAFIA,RX CONVENCIONAL}	=>	{URGÊNCIA GERAL}	0.052	0.469	1.910	44526
79	{PATOLOGIA CLINICA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{URGÊNCIA GERAL}	0.054	0.465	1.896	46202
80	{ECOGRAFIA}	=>	{RX CONVENCIONAL}	0.111	0.461	1.473	94979
81	{MEDICINA GERAL E FAMILIAR}	=>	{RX CONVENCIONAL}	0.060	0.453	1.447	51099
82	{ECOGRAFIA,PATOLOGIA CLINICA}	=>	{URGÊNCIA GERAL}	0.063	0.446	1.818	54055
83	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{ECOGRAFIA}	0.080	0.445	1.855	68295
84	{PATOLOGIA CLINICA,SERVIÇOS E TÉCNICAS GERAIS}	=>	{ECOGRAFIA}	0.056	0.443	1.845	47971
85	{PATOLOGIA CLINICA,SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.051	0.440	2.172	43709

86	{PATOLOGIA CLINICA,URGÊNCIA GERAL}	=>	{ECOGRAFIA}	0.063	0.440	1.834	54055
87	{RX CONVENCIONAL}	=>	{URGÊNCIA GERAL}	0.137	0.439	1.790	118043
88	{MEDICINA GERAL E FAMILIAR}	=>	{ECOGRAFIA}	0.058	0.439	1.830	49550
89	{ECOGRAFIA,PATOLOGIA CLINICA}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.061	0.432	2.419	52333
90	{ECOGRAFIA,PATOLOGIA CLINICA}	=>	{ANATOMIA PATOLÓGICA}	0.060	0.424	2.761	51318
91	{PATOLOGIA CLINICA}	=>	{URGÊNCIA GERAL}	0.143	0.416	1.694	122823
92	{PATOLOGIA CLINICA}	=>	{ECOGRAFIA}	0.141	0.410	1.708	121109
93	{ANATOMIA PATOLÓGICA}	=>	{ENDOSCOPIA}	0.063	0.408	4.038	53812
94	{PATOLOGIA CLINICA,SERVIÇOS E TÉCNICAS GERAIS}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.051	0.403	2.258	43709
95	{DERMATOLOGIA}	=>	{PATOLOGIA CLINICA}	0.060	0.403	1.171	51617
96	{ECOGRAFIA,PATOLOGIA CLINICA}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.056	0.396	1.954	47971
97	{OTORRINOLARINGOLOGIA}	=>	{PATOLOGIA CLINICA}	0.057	0.392	1.141	48644
98	{RX CONVENCIONAL,URGÊNCIA GERAL}	=>	{ECOGRAFIA}	0.052	0.377	1.572	44526
99	{PATOLOGIA CLINICA,URGÊNCIA GERAL}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.054	0.376	2.106	46202
100	{ECOGRAFIA,PATOLOGIA CLINICA}	=>	{GINECOLOGIA-OBSTETRÍCIA}	0.052	0.370	3.196	44804
101	{PATOLOGIA CLINICA}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.126	0.367	1.809	108358
102	{ECOGRAFIA}	=>	{ANATOMIA PATOLÓGICA}	0.086	0.357	2.326	73563
103	{ORTOPEDIA}	=>	{PATOLOGIA CLINICA}	0.060	0.356	1.035	51304
104	{RX CONVENCIONAL}	=>	{ECOGRAFIA}	0.111	0.353	1.473	94979
105	{ANATOMIA PATOLÓGICA}	=>	{RX CONVENCIONAL}	0.054	0.351	1.121	46235
106	{OFTALMOLOGIA}	=>	{PATOLOGIA CLINICA}	0.061	0.350	1.019	52747
107	{ANATOMIA PATOLÓGICA}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.054	0.350	1.726	46114
108	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{URGÊNCIA GERAL}	0.062	0.350	1.425	53637
109	{RX CONVENCIONAL}	=>	{ORTOPEDIA}	0.108	0.344	2.049	92394
110	{SERVIÇOS E TÉCNICAS GERAIS}	=>	{ECOGRAFIA}	0.068	0.337	1.402	58609
111	{PATOLOGIA CLINICA}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.116	0.336	1.881	99265
112	{ECOGRAFIA}	=>	{GINECOLOGIA-OBSTETRÍCIA}	0.080	0.332	2.866	68386
113	{ECOGRAFIA}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.080	0.331	1.855	68295
114	{RX CONVENCIONAL}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.104	0.331	1.634	89020
115	{RX CONVENCIONAL}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.103	0.329	1.840	88340
116	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.058	0.327	1.612	50135
117	{ECOGRAFIA}	=>	{URGÊNCIA GERAL}	0.078	0.325	1.323	66916
118	{URGÊNCIA GERAL}	=>	{ECOGRAFIA}	0.078	0.317	1.323	66916
119	{ORTOPEDIA}	=>	{ECOGRAFIA}	0.050	0.298	1.242	42950
120	{OFTALMOLOGIA}	=>	{RX CONVENCIONAL}	0.052	0.296	0.947	44627
121	{SERVIÇOS E TÉCNICAS GERAIS}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.058	0.288	1.612	50135
122	{ECOGRAFIA}	=>	{SERVIÇOS E TÉCNICAS GERAIS}	0.068	0.284	1.402	58609
123	{ECOGRAFIA}	=>	{IMAGIOLOGIA MAMÁRIA}	0.064	0.268	2.878	55212
124	{SERVIÇOS E TÉCNICAS GERAIS}	=>	{ANATOMIA PATOLÓGICA}	0.054	0.265	1.726	46114
125	{URGÊNCIA GERAL}	=>	{SERVIÇOS ESPECIAIS CARDIOLOGIA}	0.062	0.254	1.425	53637
126	{SERVIÇOS E TÉCNICAS GERAIS}	=>	{EQUIPA CIRURGICA}	0.052	0.254	2.507	44249
127	{PATOLOGIA CLINICA}	=>	{ANATOMIA PATOLÓGICA}	0.086	0.249	1.620	73461
128	{ECOGRAFIA}	=>	{MEDICINA GERAL E FAMILIAR}	0.058	0.240	1.830	49550
129	{PATOLOGIA CLINICA}	=>	{MEDICINA GERAL E FAMILIAR}	0.080	0.233	1.774	68853
130	{URGÊNCIA GERAL}	=>	{TAC}	0.055	0.225	1.987	47391
131	{ECOGRAFIA}	=>	{ORTOPEDIA}	0.050	0.208	1.242	42950

132	{RX CONVENCIONAL}	=>	{TAC}	0.065	0.207	1.828	55585
133	{PATOLOGIA CLINICA}	=>	{TAC}	0.069	0.200	1.770	59169
134	{PATOLOGIA CLINICA}	=>	{GINECOLOGIA-OBSTETRÍCIA}	0.067	0.193	1.670	57110
135	{RX CONVENCIONAL}	=>	{MEDICINA GERAL E FAMILIAR}	0.060	0.190	1.447	51099
136	{PATOLOGIA CLINICA}	=>	{OFTALMOLOGIA}	0.061	0.179	1.019	52747
137	{PATOLOGIA CLINICA}	=>	{DERMATOLOGIA}	0.060	0.175	1.171	51617
138	{PATOLOGIA CLINICA}	=>	{ORTOPEDIA}	0.060	0.174	1.035	51304
139	{PATOLOGIA CLINICA}	=>	{EQUIPA CIRURGICA}	0.059	0.172	1.699	50882
140	{RX CONVENCIONAL}	=>	{ANATOMIA PATOLÓGICA}	0.054	0.172	1.121	46235
141	{RX CONVENCIONAL}	=>	{OFTALMOLOGIA}	0.052	0.166	0.947	44627
142	{PATOLOGIA CLINICA}	=>	{OTORRINOLARINGOLOGIA}	0.057	0.165	1.141	48644
143	{PATOLOGIA CLINICA}	=>	{Internamento GDHS}	0.055	0.160	2.023	47194
144	{PATOLOGIA CLINICA}	=>	{IMAGIOLOGIA MAMÁRIA}	0.052	0.151	1.626	44709
145	{PATOLOGIA CLINICA}	=>	{ENDOSCOPIA}	0.051	0.149	1.478	44154

Table 13 – Association Rules Complete

8.2. PAGERANK TABLE

	Node	PageRank
1	PATOLOGIA CLINICA	7.30
2	SERVIÇOS E TÉCNICAS GERAIS	5.00
3	ORTOPEDIA	4.44
4	ECOGRAFIA	4.35
5	ANATOMIA PATOLÓGICA	3.87
6	RX CONVENCIONAL	3.75
7	GINECOLOGIA-OBSTETRÍCIA	2.98
8	OTORRINOLARINGOLOGIA	2.96
9	SERVIÇOS ESPECIAIS CARDIOLOGIA	2.75
10	MEDICINA GERAL E FAMILIAR	2.67
11	TRATAMENTO MEDICINA FISICA E REABILITAÇÃO	2.12
12	TAC	1.95
13	OFTALMOLOGIA	1.91
14	DERMATOLOGIA	1.88
15	URGÊNCIA GERAL	1.75
16	ENDOSCOPIA	1.72
17	EQUIPA CIRURGICA	1.68
18	GASTROENTEROLOGIA	1.54
19	IMAGIOLOGIA MAMÁRIA	1.52
20	RESSONÂNCIA	1.43
21	UROLOGIA	1.41
22	MEDICINA FÍSICA E REABILITAÇÃO	1.39
23	CIRURGIA GERAL	1.22
24	PEDIATRIA	1.22
25	SERVIÇOS ESPECIAIS OTORRINOLARINGOLOGIA	1.19
26	MEDICINA INTERNA	1.13

27	URGÊNCIA PEDIATRIA	1.09
28	SERVIÇOS ESPECIAIS OFTALMOLOGIA	1.06
29	Internamento GDHS	1.06
30	CARDIOLOGIA	1.01
31	IMUNO-ALERGOLOGIA	0.88
32	NEURO-CIRURGIA	0.81
33	NEUROLOGIA	0.79
34	BLOCO OPERATORIO CENTRAL	0.79
35	RECOBRO	0.78
36	CIRURGIA VASCULAR	0.74
37	BLOCO CIRURGIA AMBULATÓRIO	0.72
38	SERVIÇOS ESPECIAIS MEDICINA DENTÁRIA	0.71
39	SERVIÇOS ESPECIAIS PNEUMOLOGIA	0.71
40	PNEUMOLOGIA	0.66
41	PSIQUIATRIA	0.63
42	BIÓPSIAS	0.62
43	CIR. PLÁSTICA E RECONSTRUTIVA	0.61
44	ENDOCRINOLOGIA	0.61
45	ESTRUTURAS DENTO-ALVEOLARES	0.58
46	ANESTESIOLOGIA	0.54
47	MEDICINA DENTARIA	0.52
48	PSICOLOGIA	0.51
49	ONCOLOGIA	0.49
50	PELE, ANEXOS E PARTES MOLES	0.47
51	REUMATOLOGIA	0.45
52	SISTEMA MUSCULO ESQUELÉTICO	0.42
53	SERVIÇOS ESPECIAIS NEUROFISIOLOGIA	0.42
54	SERVIÇOS ESPECIAIS IMUNOALERGOLOGIA	0.39
55	SERVIÇOS ESPECIAIS OBSTETRICIA	0.38
56	SERVIÇOS ESPECIAIS GASTROENTEROLOGIA	0.38
57	SERVIÇOS ESPECIAIS UROLOGIA	0.37
58	SERVIÇOS ESPECIAIS DERMATOLOGIA	0.35
59	INJECÇÕES PARA TERAPÊUTICAS ESPECIAIS	0.34
60	RADIOTERAPIA	0.34
61	EXAMES ESPECIAIS RADIOTERAPIA	0.32
62	CIRURGIA PEDIÁTRICA	0.31
63	SERVIÇOS ESPECIAIS PEDIATRIA	0.31
64	APARELHO DIGESTIVO	0.31
65	OLHOS E ANEXOS OCULARES	0.30
66	GRAVIDEZ E PARTO	0.30
67	SERVIÇOS ESPECIAIS GINECOLOGIA	0.29
68	SERVIÇOS ESPECIAIS CIRURGIA VASCULAR	0.27
69	NUTRIÇÃO	0.26
70	URGÊNCIA OBSTETRICIA	0.24
71	IMUNOHEMATERAPIA	0.23
72	BLOCO DE PARTOS	0.23

73	TERAPIA DA FALA	0.23
74	SERVIÇOS ESPECIAIS REUMATOLOGIA	0.21
75	CIRURGIA MAXILO-FACIAL	0.21
76	DENSITOMETRIA	0.20
77	DIÁRIAS	0.19
78	MEDICINA NUCLEAR	0.19
79	SERVIÇOS ESPECIAIS MAXILO-FACIAL	0.18
80	SERVIÇOS ESPECIAIS ORTOPEDIA	0.18
81	CHECK-UP - CONSULTAS	0.18
82	APARELHO GENITAL FEMININO	0.17
83	PEDOPSIQUIATRIA	0.17
84	DIAGNÓSTICO E TERAPÊUTICA VASCULAR	0.17
85	APARELHO RESPIRATÓRIO	0.17
86	URGÊNCIA ORTOPEDIA	0.17
87	NEFROLOGIA	0.17
88	APARELHO URINÁRIO	0.15
89	SISTEMA AUDITIVO	0.15
90	CHECK-UP - OUTROS MCDTS	0.15
91	ANESTESIA	0.15
92	URGÊNCIA OTORRINOLARINGOLOGIA	0.15
93	SISTEMA CARDIO-VASCULAR	0.14
94	APARELHO GENITAL MASCULINO	0.14
95	CHECK-UP	0.14
96	RADIOCIRURGIA	0.13
97	TEMP	0.13
98	URGÊNCIA OFTALMOLOGIA	0.13
99	HOSPITAL DE DIA	0.13
100	CIRURGIA CARDIO-TORACICA	0.13
101	ECOCARDIACA	0.12
102	NEUROPSICOLOGIA	0.12
103	SISTEMA NERVOSO	0.12
104	SERVIÇOS ESPECIAIS GAMMA KNIFE	0.12
105	MAMA	0.11
106	URGÊNCIA CIRURGIA GERAL	0.11
107	BIÓPSIA INCISIONAL, REQUERENDO SUTURA	0.11
108	ECO DOPPLER	0.11
109	INFECCIOLOGIA	0.11
110	URGÊNCIA DERMATOLOGIA	0.11
111	URGÊNCIA NEUROLOGIA	0.11
112	RX DIGESTIVOS	0.11
113	MAMOGRAFIA	0.10
114	URGÊNCIA CIRURGIA PEDIÁTRICA	0.10
115	URGÊNCIA GINECOLOGIA-OBSTETRÍCIA	0.10
116	SISTEMA ENDÓCRINO	0.10
117	URGÊNCIA MEDICINA INTERNA	0.10
118	SERVIÇOS ESPECIAIS NEFROLOGIA	0.10

119	RX URINÁRIOS	0.10
120	N/A	0.10
121	GASTROENTEROLOGIA - PF	0.10
122	SERVIÇOS ESPECIAIS PSIQUIATRIA	0.10
123	URGÊNCIA UROLOGIA	0.10
124	RUBRICAS	0.10
125	URGÊNCIA CIR. PLÁSTICA E RECONSTRUTIVA	0.10
126	ANGIOGRAFIA	0.10
127	PEQUENA CIRURGIA - PF	0.10
128	URGÊNCIA CIRURGIA MAXILO-FACIAL	0.10
129	URGÊNCIA GASTROENTEROLOGIA	0.10
130	MEDICINA DESPORTIVA	0.10
131	CHECK-UP - PATOLOGIA CLINICA	0.10
132	ADSE CIRURGIAS	0.10
133	URGÊNCIA CARDIOLOGIA	0.10
134	SISTEMA HEMOLINFOPOIÉTICO	0.10
135	ECOGRAFIA OBSTETRICA	0.10
136	URGÊNCIA PSIQUIATRIA	0.10
137	GENETICA	0.10
138	EDUC. ESPECIAL/PSICOMOTRICIDADE	0.10
139	MEDIASTINO E DIAFRAGMA	0.10
140	URGÊNCIA PNEUMOLOGIA	0.10
141	CIRURGIA TORÁCICA	0.10
142	URGÊNCIA CIRURGIA VASCULAR	0.10
143	SERVIÇOS ESPECIAIS NEUROPSICOLOGIA	0.10
144	PACOTES - CIRURGIA	0.10
145	URGÊNCIA IMUNO-ALERGOLOGIA	0.10
146	URGÊNCIA NEURO-CIRURGIA	0.10
147	KITS AMBULATORIO	0.10
148	SERVIÇOS ESPECIAIS HEMATO-ONCOLOGIA	0.10
149	REABILITAÇÃO COGNITIVA	0.10
150	URGÊNCIA ONCOLOGIA	0.10
151	URGÊNCIA REUMATOLOGIA	0.10
152	CIRURGIA CARDIACA	0.10
153	MEDICINA ANTI-ENVELHECIMENTO	0.10
154	MEDICINA E CIRURGIA ESTÉTICA	0.10
155	URGÊNCIA ANESTESIOLOGIA	0.10

Table 14 – PageRank Complete