XXVII Congreso de la Asociación Latinoamericana de Sociología. VIII Jornadas de Sociología de la Universidad de Buenos Aires. Asociación Latinoamericana de Sociología, Buenos Aires, 2009.

Mineração de dados numéricos e de conteúdos de texto. Como apoio a descobertas de conhecimento para políticas de prevenção da violência: uma pesquisa experimental.

Gilson Lima y André Silva.

Cita:

Gilson Lima y André Silva (2009). *Mineração de dados numéricos e de conteúdos de texto. Como apoio a descobertas de conhecimento para políticas de prevenção da violência: uma pesquisa experimental. XXVII Congreso de la Asociación Latinoamericana de Sociología. VIII Jornadas de Sociología de la Universidad de Buenos Aires. Asociación Latinoamericana de Sociología, Buenos Aires.*

Dirección estable: https://www.aacademica.org/000-062/294

Acta Académica es un proyecto académico sin fines de lucro enmarcado en la iniciativa de acceso abierto*. Acta Académica fue creado para facilitar a investigadores de todo el mundo el compartir su producción académica. Para crear un perfil gratuitamente o acceder a otros trabajos visite: https://www.aacademica.org*.

Mineração de dados numéricos e de conteúdos de texto

Como apoio a descobertas de conhecimento para políticas de prevenção da violência: uma pesquisa experimental[1]

Dr. Gilson Lima[2]

Ms. André Silva[3]

"A imobilidade me faz pensar em grandes espaços
onde acontecem movimentos que não tem fim."
Joan Miró

1 Esse artigo é produto dos esforços de uma pesquisa financiada pelo FINEP e que envolve uma parceria entre a Pós-Graduação do Centro Universitário IPA, o núcleo de estudos de violência e cidadania do Programa de Pós-Graduação em Sociologia da Universidade Federal do Rio Grande do Sul (UFRGS) e a Secretaria Estadual de Segurança Pública do Rio Grande do Sul através do seu Departamento de Estratégia Operacional, Divisão de Estatística.

2 **Gilson Lima**. Doutor em Sociologia pela Universidade Federal do Rio Grande do Sul – UFRGS. Professor e pesquisador da Pós-Graduação na Rede Metodista de Educação do Sul – IPA – em Porto Alegre, RS. Pesquisador do LaDCIS – Laboratório de Difusão de Ciência, Tecnologia e Inovação Social. Colaborador e Membro do Núcleo Violência y Cidadania com ênfase em metodologias informacionais –– Universidade Federal de Rio Grande do Sul. Pesquisador da Rede Nanosoma – nanociência, nanotecnologia e sociedade. E-mail: gilima@gmail.com.

3 Ms. **André Silva.** Mestre em Engenharia da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul. Pesquisador e Professor do Centro Universitário Metodista – IPA no curso Engenharia de Computação. Pesquisador do Diretório do CNPq – Ciência, tecnologia e inclusão (NITAS – Núcleo Interdisciplinar de Simbiogênese e Tecnologia Assistiva) na linha de pesquisa *Cérebro, corpo, máquinas e softwares* (tecnologia assistiva e simbiogênese). E-mail: rs.andre@gmail.com

1. INTRODUÇÃO: metodologia da mineração de dados, textos, imagens e significados

O campo de pesquisa denominado mineração de dados **(*Data Mining*)** (LOH, 2001) corresponde a um método computacional que possibilita a descoberta de conhecimentos através de procedimentos recursivos e relacionais envolvendo grandes volumes de dados numéricos, caracteres ou imagens.

Nessa pesquisa queremos também complementar a abordagem da mineração de dados com a de centros de atividades informacionais, conforme veremos a seguir.

2. GRAFOS E A ABORDAGEM DOS CENTROS DE ATIVIDADES INFORMACIONAIS. As redes têm centros: são centros simbióticos de atividades informacionais

Diante da emergência das grandes redes digitais de informação, sendo a mais fulminante delas a Internet-Web, uma das questões mais importantes que fica para respondermos é: mesmo que a grande maioria dos processos de dados das redes digitais sejam procedimentos de agregações aleatórias, tais processos permitem ou não constituir centros de atividades informacionais que podem fazer uma enorme diferença para a descoberta de conhecimento nas redes?

Há muitos anos, e mais intensamente nos últimos cinqüenta anos, matemáticos estão discutindo se as redes se formam por agregações aleatórias ou se elas se constituem totalmente ao acaso. Na Matemática, por exemplo, os estudiosos que privilegiaram o acaso na estruturação das redes aleatórias criaram belíssimas fórmulas, pois seus interesses estavam mais voltados à expressão da beleza da Matemática do que à obtenção de uma compreensão profunda das estruturações das redes.

Alguns estudiosos, também no mesmo caminho desses matemáticos, encantados com as agregações aleatórias das redes, estudaram e buscaram suas expressões e manifestações em fenômenos sociais e na natureza. É o que nos aponta Steven Johnson em *Emergência: a dinâmica de rede em formigas, cérebros, cidades e softwares* (JOHNSON, 2003).

Johnson segue toda a trajetória do seu livro encantado com a complexidade do modo aleatório da estruturação emergente das redes complexas, que reafirmam a tese da organização emergente, segundo a qual a beleza da auto-organização é produto de uma complexidade tipo *botton-up*, ou seja, agentes individuais que residem em uma escala baixa começam a produzir comportamentos que irão residir em uma escala acima deles: seja formigas que criam colônias, seja cidadãos que criam comunidades e cidades, seja *softwares* que criam recursos de apoio cognitivo aos seus usuários.

Pensamos diferente: as redes têm centro, um centro de atividades informacionais e sociológicas, e queremos demonstrar aqui o potencial do conceito de Centros de Atividades, também conhecidos como *conectores,* a partir de uma abordagem de redes que pode ser *uma diferença que faz a diferença* em uma política pública de prevenção à violência.

De um modo sintético podemos resumir que estamos opondo à ideia de emergência simples, que tem uma característica espontânea e aleatória, uma abordagem (centro de atividade) que tenta *descobrir na estruturação relacional de uma rede de eventos os padrões chaves que compõem o núcleo central da cadeia de relações entre dados e eventos de uma rede*. Esses padrões são constituídos pelos próprios encadeamentos de relacionamentos nas redes.

Dito de outro modo, o centro de atividade é tão vital para a rede de eventos que mexer nele, extrair dele qualquer intervenção implica em alterar profundamente ou até mesmo desconstruir totalmente toda a rede de eventos constituída. Mexer em outros pontos e conectores não centrais da rede apenas afeta dinâmicas parciais da rede.

Não pensamos no tradicional conceito de variáveis (tipo linhas relacionando com colunas) com indicadores de presença e ausência das variáveis para verificar as significações estatísticas hierarquicamente relacionadas - típicas de um padrão estatístico tradicional, mas de padrões relacionais que se constituem em núcleos centrais que somente podem ser identificados pelas suas teias de conectores, conectores do tipo tudo se relaciona com tudo, mas não na mesma intensidade, força e poder de estruturação.

Barabãsi resume essa questão em uma grande lição: se até o século XX vivemos uma era de descobertas, relacionadas à forma como entendemos e usamos as propriedades individuais de objetos tão diferentes como moléculas, aviões e sites, o século XXI está revelando que será o que permitirá estudarmos e descobrirmos como as propriedades individuais de todos esses objetos e fenômenos se relacionam (BARABÃSI, 2002a).

Pensamos, então, que as redes têm centros, que são centros significativos de atividades informacionais. Encontrá-los pode ser a importante e significativa diferença que faz toda a diferença. Certa vez, o filósofo Gregory Bateson afirmou que informação

não é dado, definindo informação como a menor "diferença que faz a diferença" (HILLIS, 2000, p. 12). Perguntaríamos, então, onde residiria a diferença que faz a diferença para a prevenção da violência?

Pensamos em uma nova abordagem da *Teoria de redes,* com base em grafos que apontam para a constituição, no interior dessas redes, de *Centros de Atividades*. Para citar um exemplo, mesmo uma rede de dados muito complexa e aparentemente caótica, como a *World Wide Web*, tem seus centros de atividades, e assim é uma rede capaz de absorver facilmente falhas aleatórias (como um site que sai do ar), mas está fadada ao desastre se tiver de enfrentar um ataque dirigido.

Vejamos, então, o que seria um centro de atividades em agregações complexas de redes e o recurso metodológico proporcionado pela matemática de grafos.

Grafos são redes que consistem em *nós* conectados por arestas ou arcos. Em grafos direcionados, as conexões entre *nós* são direcionais e chamadas de arcos. Em grafos não-direcionais, as conexões chamam-se arestas. Aqui estamos falando, principalmente, de grafos direcionados.

Se realizarmos uma simulação em um computador sobre os *links* da *Web*, veremos que alguns poucos *sites* (como *Amazon, Yahoo* e *eBay*) funcionam como centros de atividade. Encontraremos milhares de outras páginas da Internet apontando para eles e milhares de pessoas tentando acessar esses *sites* ao mesmo tempo (LIMA, 2005a: 252).

Na literatura vamos encontrar como uma das primeiras possibilidades de formalização matemática do fenômeno da teoria das redes o conceito de grafos introduzido pelo matemático Leonhard Euler já no século XVII (HARARY, 1972).

Em um estudo de relacionamentos em redes Barabãsi demonstrou que as redes têm-se constituído em um fenômeno que se dá como se o mundo fosse pequeno. Segundo cálculos de Albert-Lászlo Barabãsi, uma página da *Web* está a somente 19 cliques de qualquer outra, ainda que uma esteja sediada no Japão e a outra em Honduras. A explicação para o fenômeno é simples. Preferimos conectar-nos a quem já é conectado. Páginas da *Web* com muitos *links* têm uma chance maior de receberem ainda mais *links*, pois já são conhecidas. Para tanto, um outro termo importante para a teoria dos grafos é o de **C*luster,*** *que é* um agrupamento ou subconjunto de atratores (BARABÃSI, 2002b: 36).

Para Castells uma rede pode ser definida como um conjunto de nós interconectados (CASTELLS, 1999). Tais nós podem ser pessoas, grupos ou outras unidades e as interconexões são relações, conjuntos de laços que respeitam um mesmo critério de relacionamento, dado um conjunto de nós. No entanto, como lembrou Deleuze, ao invés de pensarmos que os objetos estão se movendo através do espaço com identidade permanente, como no ponto de vista mecanicista, entendemos que tudo está basicamente se desdobrando (DELEUZE, 1991: 22).

O que importa aqui é chamar a atenção para a convergência de informações matemáticas, onde a dobra nos remete ao dobro, à bifurcação e também a um processo exponencial e dissipativo. A ideia de desdobramento tem sua aparição na matemática de Leibniz, através dos conceitos de cálculo diferencial e de mônada.

Deleuze joga com a palavra latina *plica* para a dobra, pois toda dobra provém de uma *explicação.* Assim, podemos pensar também que um conhecimento *implica* algo quando ele se dobra em articulações endógenas e que ele *explica* alguma coisa quando se desdobra e transporta seu sentido para outro(s) conhecimento(s). Explicar-implicar-complicar formam a tríade da dobra, de acordo com as variações da relação Uno-múltiplo (DELEUZE, 1991: 42-43).

Geralmente a expressão rede é aqui tratada, não como uma rede social convencional, mas como uma rede de dados ou de eventos, ou seja, um tipo específico de rede em que os nós ou atratores não são de pessoas ou grupos, mas de uma população de dados em rede. Nesse caso, os nós interconetados (os nódulos) possuem e expressam um desdobramento (dobra): a dualidade estrutural (GIDDENS, 1978, 1989, 1999). Giddens chama esse processo de *dupla hermenêutica.*

Na perspectiva da dupla hermenêutica Giddens defende que a reflexividade na modernidade informacional ocorre por intermédio e existência de uma "hermenêutica dupla", em que o primeiro meio de interpretação é o agente social e o segundo meio de interpretação é o sistema especialista. Neste último é possível identificar padrões que, em uma dada ordem, produzem e se reproduzem em âmbito simbólico, econômico, político e de legitimação (GIDDENS, 1978).

Não há nada, no entanto, que assegure que esses padrões serão reproduzidos da mesma maneira, pois é sempre pela posição no fluxo que podemos identificar os padrões, a sistemidade, a mobilidade dos atores nas várias situações interativas. A(s) posição(ões) é que permite(m) a construção da identidade pessoal e grupal de uma população de eventos ou dados. Essas posições e os fluxos das posições podem ser mais ou menos afetadas por tempos de normalidade interativas e por tempos de crise pessoais, institucionais e sistêmicas (GIDDENS, 1999).

Quanto maior for a tentativa de aproximação da relação estrutural entre o evento social e registros dos eventos informacionais maior será a capacidade de utilizar recursos mais adequados de análise sociais de redes de dados. É com base nesse conceito que vamos tentar analisar os homicídios.

Na abordagem que estamos chamando de Centro de Atividades busca-se encontrar no contexto da produção da rede social de dados um conjunto de atratores (ou nódulos de atração forte) muito atuantes (expresso por índices de presença). É importante estarmos atentos ainda para o fato de que um índice de ausência pode indicar também informação estruturante para um sistema (um atrator ao avesso ou o que Prigogine denominou de dissipação na estrutura) e a ausência e presença de atratores fracos e fortes são vitais nessa abordagem para constituir uma efetiva análise do fluxo estruturante da produção do evento social em questão (LIMA, 2005a).

3. A INFORMAÇÃO COMO PREVENÇÃO DA VIOLÊNCIA: Dois exemplos

Vamos apresentar dois exemplos que se refere à parte experimental de uma pesquisa que desde 2007 estamos realizando. Esse experimento começou com dados inicialmente pesquisados pelo estudante e capitão da Brigada Militar Cilon Freitas da Silva para uma monografia do Curso de Especialização do Núcleo de Violência e Cidadania da UFRGS. Nesse primeiro momento tentamos utilizar *softwares* de redes neurais que não surtiram efeito analítico razoável.

Em um segundo momento, trabalhamos com dados extraídos da Secretaria Estadual de Segurança Pública do Rio Grande do Sul através do seu Departamento de Estratégia Operacional, Divisão de Estatística que fica em Porto Alegre. Testamos os mesmos dados junto a uma nova possibilidade envolvendo o *software* WEKA, um programa de mineração de dados.

3.1 *Um centro de atividades informacionais visando a prevenção de homicídios na violenta cidade de São Leopoldo (Estado do Rio Grande do Sul)*

O objetivo da nossa pesquisa foi testar e simular o potencial da análise em base de mineração de dados para fins de subsidio às políticas de prevenção da violência. Para isso realizamos um conjunto de simulações de diagnoses analíticas e relacionais (SIDAR).

Iniciamos, então, uma simulação experimental sobre homicídios na cidade de São Leopoldo, localizada na região metropolitana da capital gaúcha. Queríamos a partir dessa cidade testarmos nossas hipóteses e o potencial da construção de cenários preventivos diante da abordagem baseada em centros de atividades.

Escolhemos São Leopoldo por se tratar de uma das cidades mais violentas em relação a esse tipo de crime e por, de algum modo, também ser uma das poucas cidades do Estado que nos últimos anos apresenta redução na taxa de homicídios. Em 2003 São Leopoldo ficou em primeiro lugar, com 82 casos. Em 2004, foram 60.

Frente aos dados de homicídios em São Leopoldo de 2004 e 2005 priorizamos no estudo piloto dezesseis (16) variáveis para posteriormente compormos o centro de atividades informacionais dos eventos de homicídios na cidade. São elas: 1. Faixa etária das vítimas; 2. Dias da semana do homicídio; 3. Horários em que ocorreram; 4. Suspeitos quando da ocorrência; 5. Arma de fogo; 6. Se a vítima tinha antecedentes criminais; 7. Situação penal da vítima; 8. Cor da pele da vítima; 9. Sexo da vítima; 10. Estado civil da vítima; 11. Óbito no local; 12. Tipo do local do crime; 13. Bairros onde ocorreram os homicídios; 14. Bairro de residência da vítima; 15. Distância aproximada entre a residência e o local do crime; 16. Nível de escolaridade das vítimas.[4]

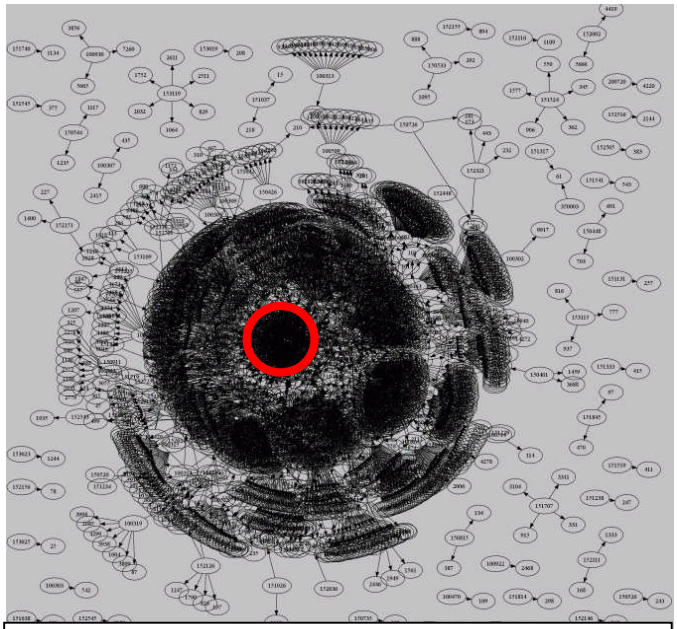


Figura 01. Representação em grafos dos fluxos da rede com atratores e nódulos relacionais

A fim de compor nossa abordagem de Centro de Atividades nesse estudo piloto encontramos dois tipos de centros:

1) **Um centro de atividades capaz de indicar um perfil das vítimas de homicídios e um centro de atividades capaz de indicar o perfil dos acontecimentos geradores dos homicídios nessa cidade**.

Vejamos:

[4] A maioria dos dados iniciais do caso da cidade de São Leopoldo aqui descritos, como já dissemos, foram extraídos inicialmente de uma orientação que realizamos a um projeto de monografia apresentado no Curso de Especialização do Núcleo de Violência e Cidadania da UFRGS pelo estudante e capitão da Brigada Militar Cilon Freitas da Silva, sob o título: *Perfil das vítimas de homicídios em São Leopoldo/RS: comparação entre duas metodologias de análise.*

Do perfil das vítimas: A maioria das vítimas de homicídios em São Leopoldo é constituída por ***homens****, negros, com idades entre 18-35 anos,* ***solteiros****,* em sua quase totalidade com, no máximo, *baixa escolaridade*, ou seja, com apenas o ensino fundamental completo, com *antecedentes criminais* e *situação penal pendente*; pessoas envolvidas em conflitos com potenciais homicidas portadores de armas de fogo e que *moram em alguns bairros da cidade:* Feitoria, Vicentina, Campina e Rio dos Sinos, que engloba a Vila dos Tocos.

Vejamos também abaixo um grafo radial Figura 02. Nesse grafo os dados que têm mais conexões vão se dirigindo ao centro da imagem. A dinâmica do grafo radial opera do seguinte modo: ele coloca no centro os elementos que estão mais conectados e na periferia os menos conectados. Isso ocorre independentemente deles estarem acima ou abaixo. Cada dado foi conectado com cada ocorrência. Todo dado que repetia na ocorrência se conectava com a ocorrência que replicava o dado e assim por diante**.** Quando mais conectores o dado possuir mais ao centro do grafo ele se dirige.

Esse processo pode-se fazer também com apenas alguns dados e não todos ou selecionando alguns dados conectores e não todos. Trata-se de um recurso altamente potente para subsidiar o processo de tomada de decisões. Por exemplo, disparando apenas as idades verificamos que entre 18 a 21 anos encontram-se 20% das vítimas de homicídio e que entre 22 e 26 anos também 20%. Podemos agora então vincular os 40% do grupo situacional de homicídio com bairros onde moram, locais da morte, antecedentes criminais (em relação ao quê?),... Vejamos um grafo radial que permite mostrar o perfil dos acontecimentos geradores (Centro de atividades).

Radial graph

FIGURA 02 – Em um grafo Radial os dados que têm mais conexões vão se dirigindo ao centro da imagem.

Do perfil dos acontecimentos geradores: As vítimas de homicídios em São

Do perfil dos acontecimentos geradores: As vítimas de homicídios em São Leopoldo morrem *nas madrugadas de sábado e de domingo*, envolvidas em fluxos de danceterias e outras festas noturnas, no *período das 18h às 06h da manhã*. Mais da metade desses acontecimentos ocorre em determinadas *vias públicas* de determinados bairros: *Rio dos Sinos (16,13%), Feitoria (14,52%), Campina (14,52%) e Vicentina (12,90%).* Os conflitos ocorrem *próximo à moradia das vítimas e nos bairros onde moram*: Feitoria, Vicentina, Campina e Rio dos Sinos (Vila dos Tocos).

3.2 Mineração de texto. Análise de descrições sobre homicídios realizadas em ocorrências policiais no Estado do Rio Grande do Sul entre 2005 e 2006

Dando continuidade a nossa pesquisa financiada pelo FINEP, fomos cruzar os dados da simulação com aqueles obtidos mediante acesso a dados brutos das ocorrências policiais de todos os homicídios no Estado do Rio Grande do Sul que nos foram fornecidos pelo Departamento de Estratégia Operacional, Divisão de Estatística da Secretaria de Segurança Pública do Estado do Rio Grande do Sul.

Ao analisarmos as planilhas verificamos, de imediato, um sistema de ocorrência muito antigo e precário, mas ao mesmo tempo pudemos constatar que uma precisa *mineração de textos* no campo livre, no qual os policiais realizavam suas descrições de campo nas ocorrências, poderia nos fornecer um enorme caldo de indicações e suportes para descoberta de conhecimento visando a constituição de um novo sistema informatizado de ocorrências no Estado.

Vejamos alguns grafos para fins demonstrativos:

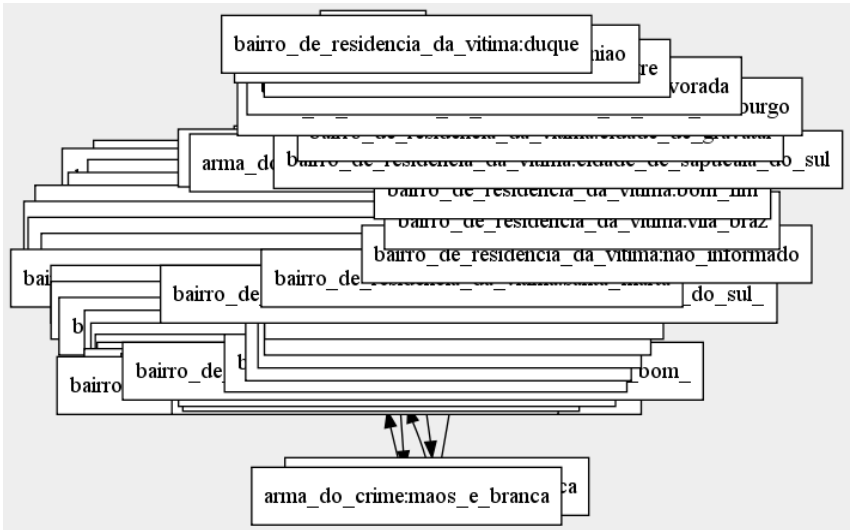
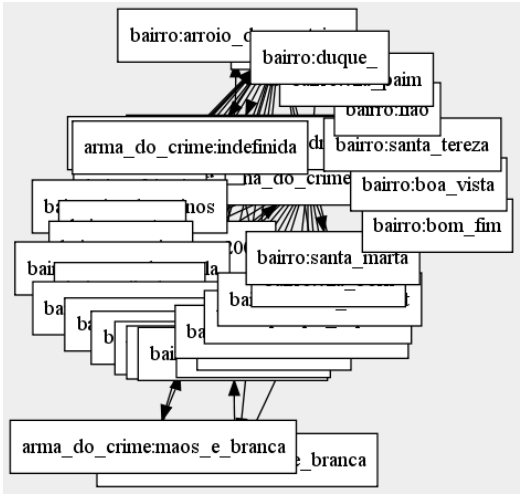


FIGURA 03 - Esse grafo nos mostra a conexão entre armas do crime, cor da pele da vítima e bairro onde ocorreu o homicídio

FIGURA 04 - Esse grafo nos mostra a conexão entre armas do crime e os bairros



Confrontando os dados iniciais de São Leopoldo, com os dados gerais de mineração de texto das ocorrências policiais de imediato compomos três fluxos centrais que nos possibilitaram gerar diagramas em uma abordagem de eventos orientados por processos. Encontramos três grandes centros de processos: 1. *Processo de identificação-localização*, 2. *Processo do evento-vítima* 3. *Processo do evento-agressor.* O primeiro objetivo aqui é fornecer um subsídio para primeiro conscientizar a autoridade pública da necessidade de constituir um novo sistema de informação envolvendo a ocorrência policial e a investigação. O segundo objetivo é fornecer os passos preliminares para a construção dos algoritmos desse novo sistema, a fim de que priorize e valorize as rotinas relacionais de entrada de dados para fins de posterior abordagem de prevenção da violência com muito mais eficácia e precisão.

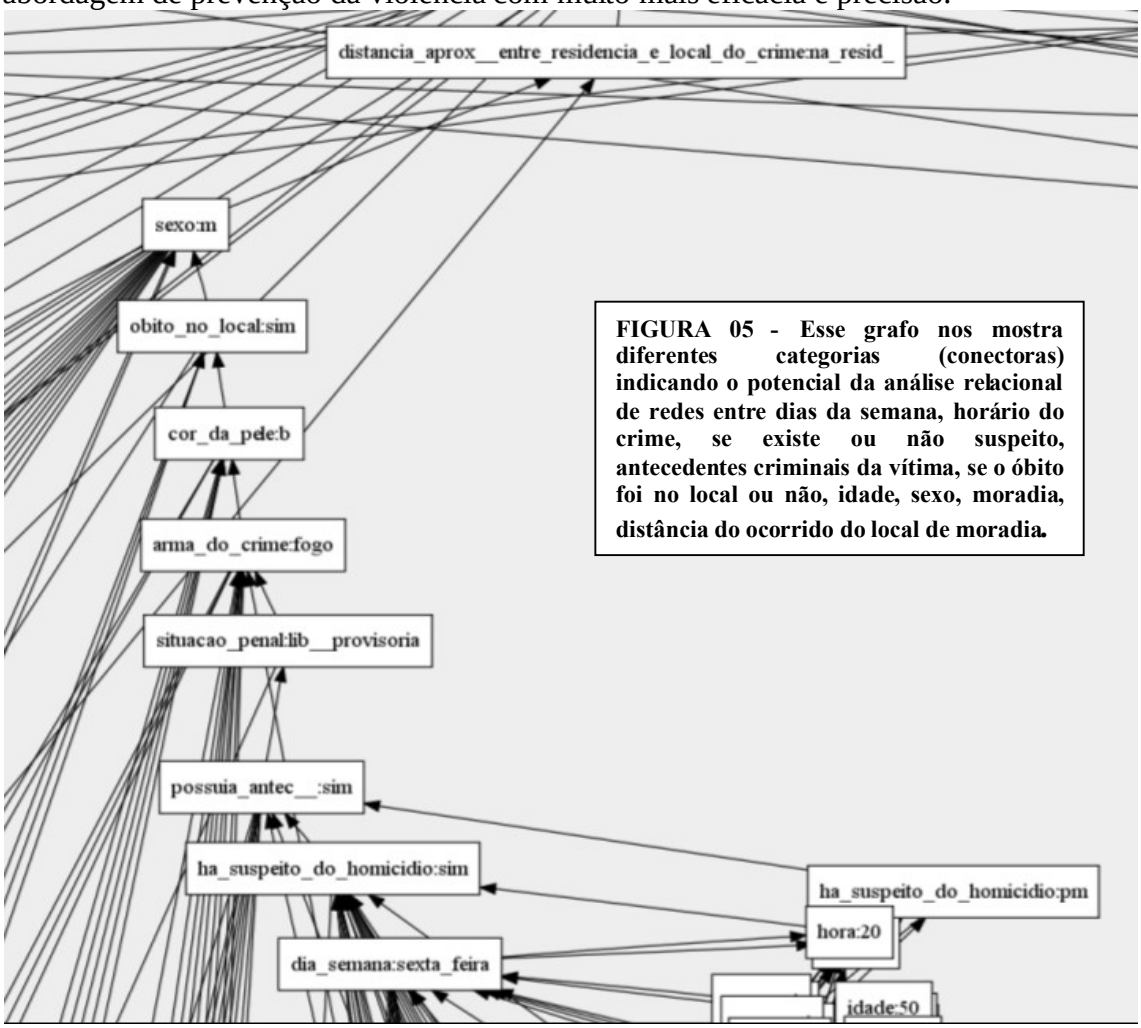


FIGURA 05 - Esse grafo nos mostra diferentes categorias (conectoras) indicando o potencial da análise relacional de redes entre dias da semana, horário do crime, se existe ou não suspeito, antecedentes criminais da vítima, se o óbito foi no local ou não, idade, sexo, moradia, distância do ocorrido do local de moradia.

No processo do EVENTO DE IDENTIFICAÇÃO-LOCALIZAÇÃO nós encontramos 23 códigos descritores. No processo do EVENTO VÍTIMA nós encontramos 30 códigos descritores, muitos desses descritores podendo ser ainda mais sub-decodificados. No processo do EVENTO AGRESSOR nós encontramos 23 códigos descritores, sendo que também muitos desses descritores poderiam ser ainda mais sub-decodificados. Vejamos, então, uma sintetização em formato de diagrama dos três grandes processos centrais encontrados e o conjunto numerado de códigos descritores em cada um deles.

DESCRITOR - EVENTO DE IDENTIFICAÇÃO-LOCALIZAÇÃO

Encontramos **23 códigos descritores**. Os códigos descritores estão em forma de número.

CAMPO DO I. D.

1) Número de identificação da ocorrência). Trata apenas do número oficial do evento.

De imediato encontramos um nó crítico.
Verificamos que o número da ocorrência inicial na Polícia Civil, não é o mesmo durante todo o processo na Justiça. No Ministério Público o evento ajuizado ganha um número diferenciado e o mesmo ocorre no processo judicial, Assim, uma tentativa de homicídio – como um acidente de trânsito doloso -, onde a vítima foi sorrida e levada com vida a um hospital pode ter-se mantido como registro de dados, uma tentativa de homicídio e não efetivamente um homicídio. Assim como um determinado agressor homicida pode em um posterior julgamento final da justiça ser considerado inocente. SUGERIMOS QUE UM SISTEMA DE OCORRÊNCIAS SEJA INTEGRADO EM TODAS AS INSTÂNCIAS, CONSTITUINDO APENAS UM ÚNICO EVENTO INFORMACIONAL DE IDENTIFICAÇÃO EM UM MESMO FLUXO RELACIONAL DE DADOS.

CAMPO DA TEMPORALIDADE

2) Ano; 3) dia; 4) mês; 5) feriado em dia de semana; 6) dia da semana; 7) mês; 8) dia do mês; 8) hora

Sugerimos decodificar no máximo o processo de parametrização da temporalidade, permitindo que o novo sistema informacional possa realizar relatórios relacionais de temporalidade muito mais detalhados. Por exemplo, vimos que os dados do evento homicídio se concentram em dias da semana e horários. Na constituição de centros de atividades informacionais, apenas essas duas relações já permitiriam dotar ações muito significativas de prevenção.

CAMPO DA ESPACIALIZAÇÃO DO EVENTO.

9) avenida; 10) rua; 11) travessa; 12) rodovia; 13) estrada; 14) distrito; 15) beco; 16) praça; 17) vila; 18) linha; 19) zona; 20) localidade; 21) CEP; 22) cidade. 23) Distrito da Delegacia da polícia civil e do Regimento do Batalhão da Brigada Militar (polícia militar).

Aqui também sugerimos decodificar ao máximo os dados de identificação espacial. Converter, localizar e relacionar diferentes dados com a localização da ocorrência pode constituir-se em um dos principais instrumentos de política de prevenção. Aqui também incluímos o código do registro do número do distrito da delegacia e do Regimento do Batalhão da Brigada Militar como instrumento significativo de manuseio de recursos humanos e materiais para políticas de prevenção da violência. Adequação da concentração de recursos humanos e materiais é um dos instrumentos mais significativos de prevenção da violência. Esses dados se alteram e permitem uma gestão mais adequada aos efetivos indicadores das ocorrências nos locais onde elas mais se concentram.

DESCRITOR - EVENTO VÍTIMA

Encontramos 30 códigos descritores, muitos desses descritores podendo ser ainda mais sub-decodificados. Os códigos descritores estão em forma de número.

CAMPO DO I. D.

1) Número de vítimas; 2) Número de vítimas fatais; 3) Número de identificação da vítima ou de cada uma das vítimas; 4) Vítimas identificadas (nome, sobrenome); 5) ID; 6) endereço; 7) CEP; 8) escolaridade (1. Ensino fundamental, 2. Ensino médio, 3. Ensino superior); 9) Profissão formal; 10) Profissão informal; 11) Desempregado; 12) Data de Nascimento; 13) Sexo; 14) Cor da pele.

CAMPO DA TIPOLOGIA DA

15) Vítima sem situação penal; 15.1) Situação penal da vítima (1. Liberdade provisória, 2. Semi-aberto, 3. Regime fechado, 4. Foragido).

16) Vítima com antecedentes criminais [Sim/Não]. 16.1) Qual? 1. Ameaça, 2. Lesão corporal, 3. Furto, 4. Porte ilegal de arma, 5. Receptação, 6. Condução sem habilitação, 7. Dano, 8. Posse de entorpecente, 9. Tráfico de entorpecentes, 10. Homicídio, 11. Estupro, 12. Jogos de azar, 13. Transtorno da ordem, 14. Falsidade ideológica, 15. Contravenção, 16. Extorsão, 17. Maus tratos.
16.2) Descrevendo características físicas (1. Cor do cabelo, 2. Comprimento do cabelo, 3. Altura, 4. Cor da vítima, 5. Peso, 6. Tipo de vestimenta – 6.1 – descrição, 7. Idade provável).

CAMPO DA TIPOLOGIA ESPACIAL DO CORPO
LOCAL DA MORTE – LOCAL DO CORPO? Sim X

17) Local da morte. 1. Em Trânsito, 2. Em socorro, 3. Hospital, 4. Via pública, 5. Ponte, 6. Praça, 7. Residência vítima, 8. Residência outra, 9. De quem. Nome proprietário, 10. Terreno, 11. Pátio, 12. Terreno baldio, 13. Chácara, 14. Casa noturna, 15. Bar, 16. Comércio (Tipos: 1. Lojas, 2. Armazéns, 3. Estética, 4. Posto de Gasolina, 5. Bar, 6. Casa Noturna).

LOCAL DO CORPO. LOCAL DA MORTE – LOCAL DO CORPO? Não. Se Diferente. Y

18) Local da morte. 1. Em Trânsito, 2. Em socorro, 3. Hospital, 4. Via pública, 5. Ponte, 6. Praça, 7. Residência vítima, 8. Residência outra, 9. De quem. Nome proprietário, 10. Terreno, 11. Pátio, 12. Terreno baldio, 13. Chácara, 14. Casa noturna, 15. Bar, 16. Comércio (Tipos: 1. Lojas, 2. Armazéns, 3. Estética, 4. Posto de Gasolina, 5. Bar, 6. Casa Noturna).

CAMPO DETALHAMENTO – CORPO – FATO GERADOR.

19) Corpo indica: (1. Homicídio, 2. Suicídio, 3. Morte pela polícia, 4. Indicativo de execução sumária [Sim/Não]).
20) Fato Gerador do Homicídio (1. Furto, 1.1. Tipo objeto, 1.2. Descrição; 2. Vingança; 3. Passional; 4. Briga; 5. Discussão).
21) Indicativo de interferência de tráfico [Sim/Não]. 1. Dívida, 2. Confronto policial, 3. Desavença familiar, 4. Briga de ponto, 22) Rebelião, 23. Estupro, 24. Jogos de azar, 25. Transtorno da ordem, 26. Falsidade ideológica, 27. Contravenção, 28. Extorsão.

29) **TIPOLOGIA RESUMIDA**. Vítima indica local do crime (Sim/Não). Em caso positivo. 1. Via pública, 2. Comércio, 3. Residência, 4. Propriedade particular, 5. No trânsito, 6. Casa noturna

Enfim, esse é o resultado que obtivemos da mineração textual depois de redirecionamos nossos esforços para uma futura situação-objetivo que envolve codificar léxicos visando uma futura montagem de fluxos e algoritmos para um futuro e novo sistema informacional de ocorrência que, além de ter registro mais precisos, possibilite uma imensa gama de simulação de cenários para monitoramento e prevenção de ações violentas.

4. PALAVRAS FINAIS

As pesquisas e a parceria que estamos fazendo com a Secretaria de Segurança Pública do Estado do Rio Grande do Sul está demonstrando um avanço considerável nessa área da gestão pública da informação como prevenção da violência, mas no tratamento informacional mais complexo verificamos que ainda temos alguns nós críticos fundamentais a enfrentar. Vejamos alguns:

1. A falta de valorização da importância da atividade de coleta de dados junto aos funcionários públicos dos diferentes órgãos responsáveis pela captura e registros das ocorrências e dos dados envolvidos nas ações violentas. Por exemplo, na Brigada Militar ainda é mais valorizado o soldado que no batalhão comumente se diz que é o que "dá tiro" contra o burocrata que realiza os registros e cadastra os dados em um determinado chamado policial. A chegada mais próxima do momento dos eventos é crucial para a precisão das investigações e para a constituição de padrões a serem analisados no futuro.
2. Na formação dos soldados e sargentos da Brigada não existe cursos específicos de capacitação na teoria e no tratamento de informação nas ocorrências policiais. Os soldados não são sequer preparados adequadamente para os registros da ocorrência.
3. O sistema de ocorrência no Estado ainda é muito antiquado, precário e altamente primário. Precisamos modernizar e integrar todo o processo de coleta, armazenamento, tratamento e recuperação dos dados em uma abordagem mais elaborada e com sistemas de banco de dados mais inteligentes e tecnologicamente muito mais complexos e com alimentação móvel.
4. Precisamos **unificar procedimentos e números das ocorrências envolvendo todos os órgãos dos diferentes poderes do executivo** (polícia civil e brigada – aqui o processo já está mais adiantado), do **Ministério Público** (no qual ainda encontramos muitas resistências) **e do Judiciário** (que opera de modo independente). O desencontro e a desintegração dos processos de coleta, registros e guarda dos dados, bem como sua despadronização dificultam, e muito, uma ação inteligente visando à posterior prevenção e um efetivo subsídio de políticas públicas preventivas. Por exemplo, uma tentativa de homicídio registrada em uma ocorrência, em que a vítima é socorrida em um hospital e falece posteriormente, se torna uma posterior ocorrência de homicídio e não uma tentativa de homicídio. Na acusação de homicídio em que, em um julgamento posterior pelo poder judiciário, ocorre a absolvição de um acusado de crime, o perfil dos dados sobre os agressores será alterado. Hoje não temos sequer como cruzar (a não ser mediante muito esforço, de modo manual e muito trabalhoso) os dados das ocorrências com as investigações do Ministério Público e do Poder Judiciário.

Diante de volumes cada vez maiores de dados disponíveis no cotidiano e diante também do fato de que tais dados envolvidos são cada vez mais tratados de algum modo por procedimentos computacionais, o que chamamos de uma sintetização digital da emergência da sociedade da informação em detrimento da sociedade industrial (produção e consumo de mercadorias), os métodos de mineração de dados são cada vez mais importantes para o apoio à descoberta de conhecimento.

A mineração de textos nos possibilitou inúmeras descobertas de conhecimento. Não teremos como apresentar aqui todos os resultados obtidos por insuficiência de espaço. Apenas queremos registrar que a mineração de texto foi altamente significativa, muito mais do que imaginávamos no início. Ela permitiu identificar o imenso potencial que o Sistema de Ocorrência possibilitaria para a descoberta de conhecimento. O sistema atual é muito antigo e precário, no entanto, mesmo assim ele nos permitiu uma precisa *mineração*

de textos no campo livre onde os policiais realizavam suas descrições de campo nas ocorrências.

Potencializarmos esse momento de coleta de registro de dados com categorias conectoras e descritores precisos poderia ajudar muito a nos fornecer um enorme caldo de indicações e suportes para a descoberta de conhecimento. Por isso indicamos um conjunto de categorias conectoras e de descritores visando à constituição de um novo sistema informatizado de ocorrências no Estado.

Por fim, a mineração de dados nos indicou que podemos com ela ter um reforço significativo para enfrentar de modo mais preciso a produção social e auto-organizada da violência e permitir que a sociedade gaúcha viva de modo mais qualificado e digno. Falta muito, mas começamos e estamos a caminho. Os primeiros passos estão sendo dados e são eles sempre os mais difíceis.

5. BIBLIOGRAFIA

BARABÃSI, Albert-László. *Linked:* the new science of networks. New York: Plume Books, 2002a.

_____. *Statistical mechanics of complex networks.* Reviews of Modern Physics, v. 74, p. 47-97, jan. 2002b.

CASTELLS, M. A sociedade em rede. Tradução de Roneide Venâncio Majer. São Paulo: Paz e Terra, 1999. (A era da informação: economia, sociedade e cultura; v. 1)

DELEUZE, Gilles. *A dobra*: Leibniz e o barroco. Campinas: Papirus, 1991.

GIDDENS, Anthony. *Novas regras do método sociológico.* Rio de Janeiro: Zahar, 1978.

_____. *As conseqüências da modernidade.* São Paulo: Unesp, 1991.

_____. & TURNER Jonathan (Org.). *Teoria social hoje.* São Paulo: Unesp, 1999.

HARARY, F. *Graph theory*. Massachusetts: Addison-Wesley, 1972. 274 p. (Addison-Wesley Series in Mathematics)

HILLIS, Daniel. *O Padrão Gravado na Pedra:* as idéias simples que fazem o computador funcionar. Rio de Janeiro: Rocco, 2000.

JOHNSON, Steve. *Emergência:* a dinâmica de rede em formigas, cérebros, cidades e softwares. Rio de Janeiro: Jorge Zahar, 2003.

LATOUR, B.; WOOLGAR, S. *A vida de laboratório*: a produção dos fatos científicos. Rio de Janeiro: Relume Dumará, 1997.

LASCH, Scott et al. *Modernização reflexiva:* política, tradição e estética na ordem moderna. São Paulo: UNESP, 1997.

LOH, Stanley; GARIN, Ramiro Saldaña. *Web Intelligence* – Inteligência Artificial para Descoberta de Conhecimento na Web. Trabalho apresentado na *V Oficina de Inteligência Artificial*, nov. 2001, Universidade Federal de Pelotas, RS.

LIMA, Gilson. *Nômades de Pedra:* Teoria da sociedade simbiogênica. Porto Alegre: Escritos, 2005a.

_____. *As redes têm centros*: uma estratégia para migração da cultura pré-digital para a simbiose de redes sociais integradas em centros de atividades sociológicas e informacionais. *Liinc em Revista*, v.1, n.2, setembro 2005b. Disponível em: http://www.liinc.ufrj.br/revista.

LOH, Stanley; GARIN, Ramiro Saldaña. *Web Intelligence* – Inteligência Artificial para Descoberta de Conhecimento na Web. Trabalho apresentado na *V Oficina de Inteligência Artificial*, nov. 2001, Universidade Federal de Pelotas, RS, p. 11-34.

MORIN, Edgar; MOIGNE, Jean-Louis. *Inteligência da Complexidade.* São Paulo: Peirópolis, 2000.
TAVARES DOS SANTOS, José Vicente. As possibilidades das metodologias informacionais nas práticas sociológicas: por um novo padrão de trabalho para os sociólogos do século XXI. *Sociologias,* n. 5, p.114-146, 2001.
SILVA, Cilon Freitas da. *Perfil das Vítimas de Homicídios em São Leopoldo/RS:* comparação entre duas metodologias de análise. Monografia de Conclusão de curso de especialização. Porto Alegre, UFRGS, 2005. Mimeografado.