DESENVOLVIMENTO DE BASE DE DADOS EM LÍNGUA PORTUGUESA SOBRE CRIMES SEXUAIS

Adriana Andrijauskas[1],Adriana Shimabukuro[2], Prof. Dr. Rodrigo Filev Maia[3]
[1,3] Departamento de Ciência da Computação, Centro Universitário FEI
[2] Departamento de Crimes Cibernéticos, Ministério Público Federal-SP
adriana.jks@gmail.com rfilev@fei.edu.br

Resumo: Uma das principais ferramentas utilizadas por predadores sexuais para buscar e aliciar vítimas é através de ferramentas de comunicação de texto, pois elas permitem o anonimato. Pesquisas baseadas em detecção de padrões para identificar predadores sexuais estão sendo desenvolvidas com uma base de diálogos em inglês. Esta iniciação tecnológica tem como objetivo a criação de uma base em língua portuguesa que possa ser utilizada em pesquisas de acordo com as definições legais definidas pelo MPF-SP.

1. Introdução

As redes de pedofilia na internet são atualmente bastante extensas, pois o meio digital permite a troca de materiais e dados de forma anônima, fazendo com que os predadores sexuais se sintam mais seguros e com menor chance de serem descobertos.

Uma das questões associadas com o aliciamento de crianças via internet é a inocência e inexperiência dos menores, associado à crença dos pais ou responsáveis, que por estar dentro de casa, a criança está segura.

Atualmente a questão de segurança de pessoas na internet vem sendo trabalhada em diversas pesquisas e por diversas instituições, pois diversos estudos claramente identificam que a ameaça representada por predadores sexuais é bastante séria, e as consequências do ataque destes indivíduos a crianças e adolescentes perpetram marcas psicológicas e sociais demasiadamente danosas e de nem sempre de recuperação satisfatória [1], [2], [3], [4], [5], [6].

No ano de 2012 a PAN (organização de especialistas em tecnologias forenses) lançou como um de seus desafios a detecção de conversas entre predadores sexuais dentre diversas conversas. A organização ofereceu uma base contendo diversas conversas, entre elas conversas culpadas (onde existe a detecção de um predador sexual e uma vítima) e conversas não culpadas (conversas sem cunho sexual entre um maior de idade e um menor de idade).

Diversos trabalhos foram realizados usando essa base, em [7] elaborou-se um algoritmo de classificação de conversas de potenciais predadores sexuais via análise de mensagens de texto. Nos resultados iniciais obtidos foi capaz de detectar parcela substancial das conversas de predadores sexuais, porém diversas perguntas foram levantadas, sendo uma delas como seria o comportamento do algoritmo de [7] e outros algoritmos utilizados na literatura com diálogos em língua portuguesa. Portanto, este trabalho visa a eloboração de uma base de dados nos moldes propostos no desafio PAN 2012 contendo conversas em língua portuguesa.

2. Metodologia

Baseando-se na estrutura do arquivo XML referente à base fornecida pelo PAN em 2012 (figura 1), foi proposto o desenvolvimento de um banco de dados relacional. Para a população do banco foi realizada uma parceria com o Ministério Público Federal, onde foi fornecido um local específico e materiais necessários para a manipulação de dados.

```
<conversations>
  ...
  <conversation id="0042762e26ed295a8576806f5548cad9">
    <message line="3">
      <author>f069dbec9ab3e090972d432db279e3eb</author>
      <time>03:20</time>
      <text>whats up?</text>
    </message>
    <message line="4">
      <author>f069dbec9ab3e090972d432db279e3eb</author>
      <time>03:21</time>
      <text>how u doing?</text>
    </message>
    ...
    <message line="10">
      <author>f069dbec9ab3e090972d432db279e3eb</author>
      <time>04:00</time>
      <text>sse you llater?</text>
    </message>
  </conversation>
  ...
  <conversation id="0209b0a30c8eced86863631ada73a530">
    <message line="3">
      <author>0042762e26ed295a8576806f5548cad9</author>
      <time>01:17</time>
      <text>and that i dont touch u</text>
    </message>
  </conversation>
  ...
<conversations>
```

Figura 1 – Estrutura do arquivo XML do PAN.

Inicialmente definiu-se a criação de três entidades e dois relacionamentos, como demonstra o MER da figura abaixo.

Figura 2 - MER

A mais completa, maior e principal entidade a ser criada será a tabela TODAS_CONVERSAS, nela será armazenada linha a linha das conversas e suas respectivas informações, sendo manipulada através de uma chave composta, referente à coluna id_conversas e à coluna linha_mensagem, garantindo assim unicidade das linhas. Para melhor identificação dos autores culpados, será criada a tabela AUTORES_CULPADOS, contendo apenas uma chave estrangeira que fará referência à coluna autor da tabela TODAS_CONVERSAS, garantindo assim que apenas dados que existam na tabela que ela referencia possam ser escritos se os mesmos forem considerados culpados. Por último será criada uma tabela com nome CONVERSAS_CULPADAS contendo uma coluna do tipo chave estrangeira que apontará para a coluna id_conversa da tabela principal e outra coluna que armazenará o número de linha que fez a conversa ser classificada como culpada.

3. Resultados

O banco de dados relacional foi criado de acordo com todos os critérios previamente definidos (estrtutura das tabelas e suas respectivas chaves), exceto pela variável "time" referente à tabela TODAS_CONVERSAS (figura 3), pois muitos arquivos coletados pelo MPF não apresentaram a variável hora em todas as mensagens. A plataforma utilizada foi a ferramenta MySQL e atualmente a base possui por volta de 80 conversas e 2000 linhas. As mesmas conversas estão armazenadas em um arquixo XML para que possam ser utilizadas na construção de outras bases de dados.

id_conversa	linha_mensagem	autor	texto
1	1	7d439b8298b548143da3c658d4dea800	Voce deita com mico na cama pelada
1	2	7a1191dcd2b31dff6e01f5c19e12b6c4	Nao posso durmi ai
1	3	7d439b8298b548143da3c658d4dea800	Nao e pra dumir e so fica com mico ate as 3 hor...
1	4	7d439b8298b548143da3c658d4dea800	Ai eu levo voce em bora
1	5	7a1191dcd2b31dff6e01f5c19e12b6c4	Tah
1	6	709916bfe16ef8cdd6102dc5453f302f	se eu nao tiver ouando eu cheoar em casa eu ti...
1	7	709916bfe16ef8cdd6102dc5453f302f	hum
1	8	709916bfe16ef8cdd6102dc5453f302f	podemos marca nosso encontro em um local on...
1	9	709916bfe16ef8cdd6102dc5453f302f	sou muito conhecido na minha igreia
1	10	709916bfe16ef8cdd6102dc5453f302f	vc nao resoondeu
1	11	709916bfe16ef8cdd6102dc5453f302f	tem aloum problema eu ser evanoelico
1	12	13f27f55ef3622f4e987aac6a57b1ce8	Oual sua igreia?
2	1	298f72d7b668828a60d48537729533cb	oi
2	2	298f72d7b668828a60d48537729533cb	eu tenhu um shord de dormi curto mas não sei ...
2	3	298f72d7b668828a60d48537729533cb	vou tentar
2	4	c1355adef586747f642be93196c7cd06	Ele fica bem curtinho?
2	5	c1355adef586747f642be93196c7cd06	Até onde oue ele fica?
2	6	298f72d7b668828a60d48537729533cb	tipo cueca box
2	7	c1355adef586747f642be93196c7cd06	Conseouiu mandar a foto?
2	8	298f72d7b668828a60d48537729533cb	não
2	9	c1355adef586747f642be93196c7cd06	O emoresário da Larissa Manoela oode falar co...
2	10	c1355adef586747f642be93196c7cd06	Estou ocupada
2	11	c1355adef586747f642be93196c7cd06	Ai vc mostra o SHORTINHO oara ele
2	12	c1355adef586747f642be93196c7cd06	Dá?
2	13	298f72d7b668828a60d48537729533cb	eu não to em casa
2	14	c1355adef586747f642be93196c7cd06	Cadê o SHORTINHO?
2	15	298f72d7b668828a60d48537729533cb	vo ver se tenho um aoui na minha vo
2	16	c1355adef586747f642be93196c7cd06	Pôde ser calcinha de dormir
2	17	c1355adef586747f642be93196c7cd06	Tem?
3	1	79196bd95ffb6a7d7d41091df457dcff	como vou advinhar
3	2	632f09c7960c1cdff4adfca7402b478a	tipo oensei o vc tinha falado o enioo de ver vid...
3	3	632f09c7960c1cdff4adfca7402b478a	ah eu oosto de ver vídeo.assistir.fazer baounca...
3	4	79196bd95ffb6a7d7d41091df457dcff	a vc esta me convidando o fazer
3	5	0c4ce3210257eeb486a41358ad96f2dd	Aor ta tô dormindo!
3	6	ff3465f4b49d366a22add4ce07b12825	Isso e bom
3	7	ff3465f4b49d366a22add4ce07b12825	Vc só deve ta de camisola e mais nada
3	8	0c4ce3210257eeb486a41358ad96f2dd	Com a luz do quarto ta apaoada :3

Figura 3 – Tabela TODAS_CONVERSAS

Contudo, esta base de dados ainda está sob sigilo e a divisão de Crimes Cibernéticos do MPF-SP está trabalhando para permitir o uso desta base em outros ambientes que não sejam apenas das forças de segurança nacionais.

4. Conclusões

Após estudada a estrutura utilizada no PAN 2012, concluiu-se que será necessário a criação de duas bases de dados, uma relacional e uma orientada a objetos. Bases de dados relacionais são melhores para a manipulação de filtros em conversas, portanto nela que serão aplicados eventuais filtros para detecção de padrões nas conversas. Já as bases de dados orientadas a objeto são mais indicadas para realizar a manipulação de palavras e detectar o peso de seus relacionamentos (importância), podendo ser realizado o estudo das palavras através de grafos.

Para a criação do banco de dados orientado a objeto, estudos de ferramentas, de estrutura e de filtros para a sua população estão sendo realizados.

5. Referências

[1] BALFE, Myles et al. Internet child sex offenders' concerns about online security and their use of identity protection technologies: a review. Child Abuse Review, 2014.

[2] BAUMGARTNER, Susanne E. et al. Does country context matter? Investigating the predictors of teen sexting across Europe. Computers in Human Behavior, v. 34, p. 157-164, 2014.

[3] BEECH, A. R., et al. "The internet and child sexual offending: A criminological review."Aggression and violent behavior n. 13 v.3, 2008.

[4] HILLMAN, Henry; HOOPER, Christopher; CHOO, Kim-Kwang Raymond. Online child exploitation: Challenges and future research directions. Computer Law & Security Review, v. 30, n. 6, p. 687-698, 2014.

[5] KLOESS, J. A., ANTHONY R. B., LEIGH, H. "Online Child Sexual Exploitation Prevalence, Process, and Offender Characteristics." *Trauma, Violence, & Abuse* n. 15 v.2, 2014.

[6] LIVINGSTONE, S.; SMITH, P. K. Annual Research Review: Harms experienced by child users of online and mobile technologies: the nature, prevalence and management of sexual and aggressive risks in the digital age. Journal of child psychology and psychiatry, v. 55, n. 6, p. 635-654, 2014.

[7] VIVIANI, G. L.; DIAS, G. R.; MARTINEZ, G. T.; FERREIRA, I. C. **Reconhecimento de Padrões em Conversas Utilizando Redes Complexas.** Trabalho de Conclusão de Curso—Centro Universitário da FEI, 2014.

Agradecimentos

À instituição Centro universitário FEI pela realização das medidas e empréstimo de equipamentos, ao Prof. Dr. Rodrigo Filev Maia pela orientação e ao Ministério Público Federal pela parceria, especialmente à chefe do departamento de Crimes Cibernéticos, Adriana Shimabukuro.

[1] Aluno de IT do Centro Universitário FEI com bolsa CNPq. Projeto com vigência de 01/17 a 12/17.