Faculdade de Tecnologia Bandeirantes - BandTec 2^o semestre de 2015

Descobrindo relações entre categorias de conteúdo em redes sociais frequentadas por adolescentes

Pós Graduação em Big Data

Prof. Fabrício Jailson Barth

1 Introdução

Interagir com amigos em redes sociais, tais como Facebook e MySpace, tem se tornado um ritual entre os adolescentes do mundo todo. Estes adolescentes, ao mesmo tempo que interagem com os seus amigos, também estão sugeitos a propagandas de diversos produtos e empresas.

É muito importante para as empresas, e até mesmo para os usuários de redes sociais, que tais propagandas sejam melhor direcionadas, ou seja, que sejam entregues para aqueles que realmente tem interesse naquele tipo de produto.

Dado o texto produzido por adolescentes em Serviços de Redes Sociais, acredita-se que podemos identificar relações entre os textos, por exemplo: 80% das pessoas que escrevem ou lêem sobre futebol também lêem ou escreve sobre basquete. O conhecimento deste tipo de relação é importante para a personalização de layouts e conteúdos, assim como a recomendação de conteúdo.

O objetivo deste estudo é identificar relações entre tipos de conteúdos produzidos e consumidos em uma rede social.

2 Método

O método utilizado por esta análise é composto pelas seguintes fases: aquisição e pré-processamento dos dados; construção do modelo, e; análise do modelo.

3 Aquisição e pré-processamento dos dados

Para esta análise, nós vamos utilizar um dataset que representa uma amostra aleatória de 30.000 estudantes de *high school* americana que possuem perfil em uma rede social[1] em 2006. O dataset pode ser encontrado em [2].

Os atributos deste dataset são: **gradyear**: ano de graduação; **gender**: sexo, masculino ou feminino; **age**: idade representada por um número real; **friends**: quantidade de amigos na rede social; **demais atributos**: basketball, football, soccer, entre outros indicam classes onde os textos das pessoas foram classificados. Textos de usuários publicados na rede social foram processados para gerar esta tabela.

A aquisição e o pré-processamento dos dados são apresentados abaixo:

```
1 snsdata <- read.csv("dataset/snsdata.csv")
2 snsdata <- snsdata[,5:40]
```

A linha 1 do código acima é utilizada para a leitura dos dados. Aqui, estamos considerando que o arquivo CSV já foi salvo na máquina local. O objetivo desta análise é encontrar relações entre as categorias dos textos e posts das pessoas, por isso, a segunda linha filtra os atributos pessoais (gradyear, gender, age, friends).

Para utilizarmos o algoritmo **Apriori** sobre este dataset, é necessário converser os valores números em categóricos. A função $map\_to\_boolean$ implementa esta conversão.

[1] Em inglês, Social Network System - SNS

[2] https://raw.githubusercontent.com/fbarth/posGraduacao/master/ExemplosClustering/data/snsdata.csv

```
map_to_boolean <- function(attr){
  attr <- ifelse(attr >= 1, TRUE, FALSE)
}
```

Para utilizar a função acima sobre o dataset analisado, basta executar a linha abaixo:

```
snsdata <- as.data.frame(sapply(snsdata, map_to_boolean))
```

O dataset $snsdata$ está no formato $data.frame$. Para utilizarmos o algoritmo $Apriori$ com este dataset, precisamos converter o dataset para o tipo $transactions$, como apresentado abaixo:

```
tr <- as(snsdata, "transactions")
```

4 Análise Descritiva e Construção do modelo

Os itens mais frequentes são apresentados na figura 1.

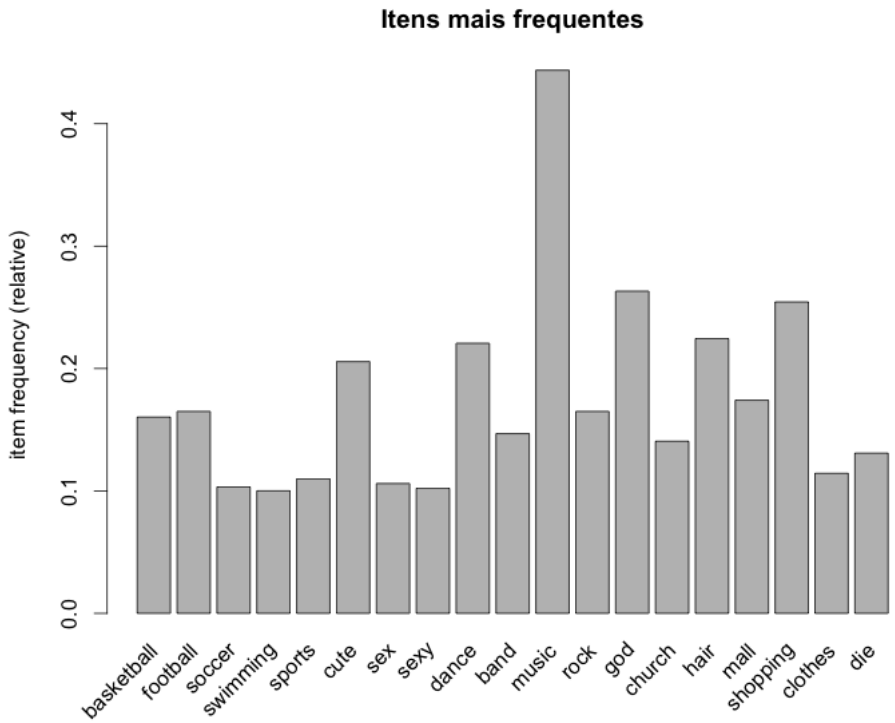

Figura 1: Itens mais frequentes encontradas nas transações

Para a construção das regras foram utilizados o valor de suporte igual a 0.01 e confiança igual a 0.6. Levando-se em consideração estes valores de suporte e confiança foram encontradas 596 regras.

Na figura 2 é possível visualizar as regras levando-se em consideração o suporte, confiança e lift de cada regra. Na figura 2 é possível visualizar que a maioria das regras tem um suporte próximo de 0.02, que existem poucas regras com confiança maior que 0.8 e que o lift das regras varia de 2 até 12. A maioria das regras esta concentrada entre suporte de 0.01 até 0.02 e confiança entre 0.6 e 0.8. Algumas das regras que se encontram entre este espaço tem lift de até 12.

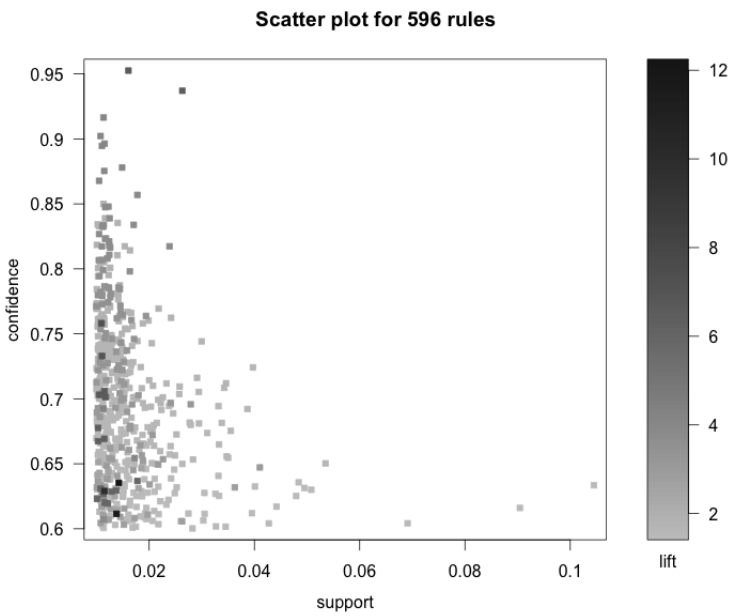

Figura 2: Distribuição das regras levando-se em consideração o suporte, confiança e lift.

Na figura 3 é possível visualizar uma matriz de antecedentes (LHS) versus consequentes (RHS) colorida levando-se em consideração os valores de confiança e lift.

Através da figura 3 é possível verificar que das 36 classes de conteúdo, apenas 10 estão do lado direito da regra. A maioria das regras possuem como consequente os itens *hair* (item 1) e *music* (item 10). A lista de itens é apresentada abaixo:

```
Itemsets in Consequent (RHS)
 [1] "{hair}" "{sex}" "{kissed}" "{band}" "{football}" "{basketball}" "{god}"
 [8] "{shopping}" "{hollister}" "{music}"
```

Os comandos utilizados para gerar as figuras 2 e 3 são apresentados abaixo:

```
plot(rules)
plot(rules, method = "matrix", measure = c("lift", "confidence"),
     control = list(reorder = TRUE))
```

Para fazer uma análise mais detalhada das regras geradas, optou-se por apresentar apenas as cinco regras com maior confiança e as cinco regras com maior lift.

As cinco regras com maior confiança foram identificadas através dos comandos:

```
rules1 <- head(sort(rules, by="confidence"), 5)
plot(rules1, method="graph", control=list(type="items"),
     main="Cinco regras com maior confiança")
```

A figura 4 apresenta de forma gráfica as regras com maior confiança encontradas e a tabela 1 apresenta a mesma informação no formato de tabela. Ao ver tanto a figura 4, como a informação contida na tabela 1 é fácil percebe que quando uma pessoa acessa conteúdos sobre *marching* ou quando uma pessoa acessa conteúdos sobre *marching* e *music* em 94% ela irá acessar conteúdo sobre *band* também. Por outro lado, combinações dos temas *sex,music,blonde,kissed,clothes* e *band* levam ao tema *hair* com uma confiança sempre maior que 90%.

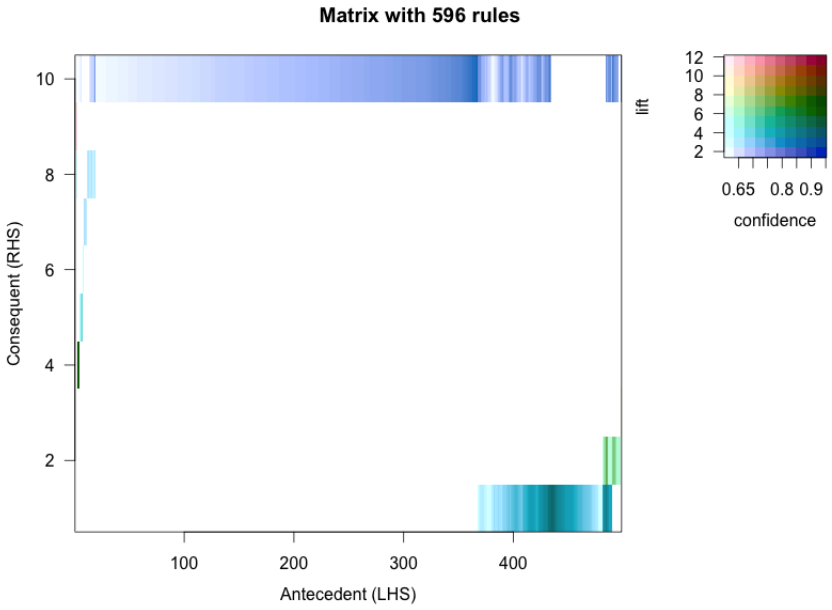

Figura 3: Distribuição das regras levando-se em consideração os antecedentes e consequentes de cada regra.

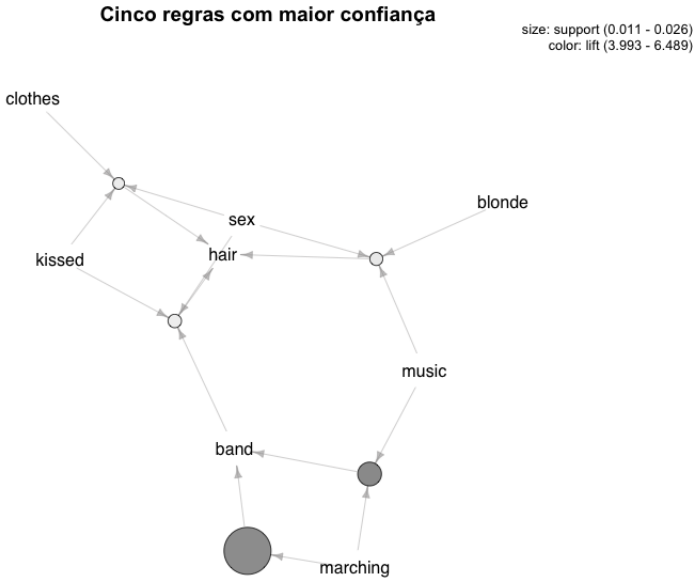

Figura 4: Representação gráfica das 5 regras identificadas com maior confiança.

	lhs	rhs	support	confidence	lift
14	{marching,music}	{band}	0.02	0.95	6.49
2	{marching}	{band}	0.03	0.94	6.38
337	{sex,music,blonde}	{hair}	0.01	0.92	4.08
351	{sex,kissed,clothes}	{hair}	0.01	0.90	4.02
357	{sex,kissed,band}	{hair}	0.01	0.90	3.99

Tabela 1: Cinco regras identificadas com maior confiança.

A figura 5 apresenta de forma gráfica as regras com maior lift encontradas e a tabela 2 apresenta a mesma informação no formato de tabela. Na figura 5 fica bem clara a existência de dois grupos de regras. Um grupo que tem como consequência *sex* e *kissed* e outro grupo que tem como consequência *hollister*.

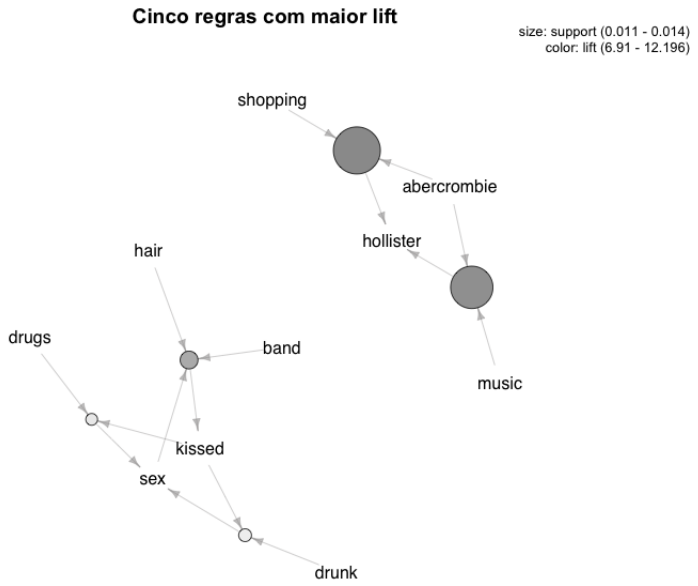

Figura 5: Representação gráfica das 5 regras identificadas com maior lift.

	lhs	rhs	support	confidence	lift
18	{shopping,abercrombie}	{hollister}	0.01	0.64	12.20
19	{music,abercrombie}	{hollister}	0.01	0.61	11.74
359	{sex,band,hair}	{kissed}	0.01	0.63	9.92
24	{kissed,drugs}	{sex}	0.01	0.76	7.15
75	{kissed,drunk}	{sex}	0.01	0.73	6.91

Tabela 2: Cinco regras identificadas com maior lift.

5 Considerações

O objetivo deste trabalho foi identificar relações entre tipos de conteúdos produzidos e consumidos em uma rede social por adolescentes. Neste estudo foram considerados uma amostra de 30.000 pessoas que acessaram conteúdo classificado em 36 tipos diferentes, entre eles: *basketball*, *football*, *soccer*, *sex*, *kissed*, entre outros.

Identificou-se que os assuntos mais populares nesta amostra são: *music* presente em mais de 40% dos perfis dos usuários, *god* e *shopping* presentes em mais de 20% dos perfis dos usuários.

Para identificar as relações foi utilizado o algoritmo *Apriori* e cada perfil de usuário foi considerado como uma transação. Foram utilizados o valor de suporte igual a 0.01 e confiança igual a 0.6. Foram encontradas 596 regras.

As regras com confiança maior que 90% são regras relacionadas a *band* e *hair*, por exemplo: quando alguém acessa material sobre *marching* e *music* então com 95% de confiança também irá acessar conteúdo sobre *band*. Quando alguém acessa material sobre *sex* e *kissed* e *blonde* então com 92% de confiança também irá acessar material sobre *hair* (tabela 1).

As regras com maior lift são regras que tem como consequência categorias com menor frequência, entre elas: *hollister*, *kissed* e *sex*. Na figura 5 e na tabela 2 fica claro a relação das categorias *kissed*, *drugs* e *drunk* com *sex* e *shopping*, *abercrombie* e *music* com *hollister*. O suporte das regras relacionadas com *hollister* tem um suporte maior que as outras regras.