FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM
MATEMÁTICA APLICADA

Dinâmica de Disseminação de Notícias em Redes Complexas

por

Elisa Mussumeci

Rio de Janeiro
2015

FUNDAÇÃO GETÚLIO
VARGAS

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM
MATEMÁTICA APLICADA

Dinâmica de Disseminação de Notícias em Redes Complexas

”Declaro ser o único autor do presente projeto de monografia que refere-se ao plano de trabalho a ser executado para continuidade da monografia e ressalto que não recorri a qualquer forma de colaboração ou auxílio de terceiros para realizá-lo a não ser nos casos e para os fins autorizados pelo professor orientador”

Elisa Mussumeci

Orientador: Flavio Codeço Coelho

Rio de Janeiro
2015

ELISA MUSSUMECI

Dinâmica de Disseminação de Notícias em Redes Complexas

"Monografia apresentada à Escola de Matemática Aplicada
como requisito parcial para obtenção do grau de Bacharel
em Matemática Aplicada"

Aprovado em ___ de _____ de _____ .
Grau atribuido ao Projeto de Monografia: ___ .

Professor Orientador: Flávio Codeço Coelho
Escola de Matemática Aplicada
Fundação Getulio Vargas

Conteúdo

1 Introdução **4**
1.1 O Projeto Media Cloud Brasil 4
1.2 Referencial Teórico 4
1.2.1 O espalhamento de notícias como um processo de contágio 4
1.2.2 Processos Epidemiológicos em Redes Complexas 5
1.2.3 Processamento de Linguagem Natural 5

2 Objetivo **5**

3 Metodologia **6**
3.1 Rede de Disseminação Real 6
3.1.1 Matriz de Documentos 6
3.1.2 Construção da Rede 8
3.2 Simulação Rede de Disseminação 9
3.2.1 Rede Completa 9
3.2.2 Simulação 10

4 Resultados **10**
4.1 Rede de Disseminação Original 10
4.2 Simulação Rede de Disseminação 10

5 Conclusão **10**

Resumo

O processo de formação de opinião é fortemente influenciado pela mídia digital. Entretanto pouco se sabe sobre o processo de disseminação de notícias e os fatores que determinam o alcance de cada notícia.

A disseminação de uma notícia se dá por meio de um ou mais caminhos em uma rede desconhecida de influência entre formadores de opinião (produtores de notícias). Este padrão pode ser recuperado, com algum grau de incerteza, a partir de dados da sequência temporal das publicações sobre um mesmo tema, e dos links nelas contidos.

Este projeto tem como objetivo caracterizar as redes de interligação de veículos de mídia e modelar a dinâmica do espalhamento de notícias, a fim de prever tendências e mapear questões de interesse.

1 Introdução

Atualmente a internet é um dos principais meios de veiculação de notícias e informação do país. Com o crescente número de pessoas aderindo à redes sociais, o compartilhamento de notícias aumentou significamente, o que tornou fundamental o papel das mídias digitais no acesso à informação.

Consideramos como mídia digital todo e qualquer veículo difusor de informação contido na internet brasileira, como jornais, revistas e blogs independentes. Cada mídia presente na internet possui sua própria periodicidade, alcance, público e credibilidade, o que afeta diretamento no processo de disseminação da informação.

Utilizando do fato que essas mídias digitais são fundamentais na disseminação da informação, e que com isso possuem um forte papel influenciador no processo de formação de opinião, podemos definir como importante entender como as notícias se formam e se espalham na internet brasileira.

Um dos métodos de se estudar a disseminação da informação é utilizar modelos epidemiológicos. (referencia de artigo) em (ano do artigo) conseguiu modelar a disseminação de (tabela de alguma coisa) no tempo através de modelos epidemiológicos, como SIR, SIS entre outros. Ese tipo de abordagem vem sendo utilizada também para entender redes de contatos em redes sociais como (incluir referencia). MELHORAR PARAGRAFO

Neste trabalho utilizaremos redes complexas e modelos epidemiológicos para modelar a disseminação de notícias na mídia brasileira. Para isso estudaremos os caminhos percorridos em uma rede de disseminação criada através de modelos de recuperação de informação utilizados em cima da base de dados do Projeto MediaCloud Brasil .

referência

1.1 O Projeto Media Cloud Brasil

Para a realização deste trabalho, foram utilizados os dados do projeto MediaCloud Brasil. O MediaCloud Brasil é um projeto concebido e mantido pelo NAMD/EMAp da Fundação Getúlio Vargas, e vem ao longo dos últimos três anos monitorando mais de cem mil veículos de mídia da internet brasileira. Possui em sua base de dados mais de 1.6 milhão de artigos capturados.

falar sobre o que e como o Media cloud captura os artigos

O projeto utiliza como banco de dados o MongoDB, um banco de dados de documentos open-source de alta performance. O MongoDB é classificado como um banco de dados 'NoSQL', uma vez que evita a tradicional estrutura baseada em tabela relacional e utiliza documentos JSON com esquemas dinâmicos para armazenamento dos documentos. A vantagem de utilizar o JSON é realizar a integração de dados em certos tipos de aplicações de forma mais fácil e mais rápida.

Que tipo de aplicações?

Falar que uma vez armazenado, o banco de artigos é indexado para permitir buscas textuais

1.2 Referencial Teórico

1.2.1 O espalhamento de notícias como um processo de contágio

Uma epidemia é caracterizada pela incidência de grande número de casos de uma doença em um curto período de tempo. Sabemos que as doenças se espalham através do contágio entre infectados, mas como podemos definir esse contágio?

Cada doença possui uma forma própria de transmissão, por exemplo, a gripe é uma doença viral e se transmite a partir de vias orais, já a dengue é transmitida a partir da picada de um mosquisto, assim como a febre amarela. Saber a forma de contágio de uma doença é fundamental para que possamos entender sua disseminação e modelar sua epidemia.

Ao observar o comportamento de notícias, assuntos e histórias na mídia, podemos ver o surgimento de memes e histórias 'virais'. Esses tipos de notícias são chamadas de virais por se espalharem muito rápido e obterem um alcançe grande na população. Algumas dessas notícias se sustentam por um longo tempo na mídia, e outras são esquecidas rapidamente.

Se pensarmos que estamos lidando com um processo de disseminação, que possui uma taxa de espalhamento e uma taxa de esquecimento, podemos facilmente fazer uma comparação à modelos epidemiológicos, principalmente ao modelo SIR, que possui uma taxa de infecção e uma taxa de recuperação. Dessa forma, podemos estudar como uma notícia se espalha da mesma forma que modelamos uma epidemia.

Para modelar a disseminação de notícias da mesma forma que modelamos a de doenças, precisamos que nosso modelo seja compatível com o de uma epidemia, ou seja, precisamos definir os infectados, suscetíveis e o método de contágio.

Em nosso modelo, uma notícia/assunto é a doença, e os infectados são todos os artigos que falam sobre essa notícia. Para definir o contágio da nossa notícia, consideramos que um artigo infecta o outro quando ele influência o outro. Ou seja, definimos que quando um artigo exerceu influência sobre um outro, ele infectou esse outro artigo.

Dessa forma, podemos ver a similaridade entre ambas modelagens, o que deixa claro a possibilidade de modelar a disseminação de notícias através de modelos epidemiológicos. O que nos falta descobrir são os parâmetros que utilizaremos para realizar essa modelagem de forma que fique o mais verídica possível.

1.2.2 Processos Epidemiológicos em Redes Complexas

1.2.3 Processamento de Linguagem Natural

explicar aqui a teoria por trás de todas as técnicas de NLP que vc usa: Tokenização, TF-IDF, etc.

representação vetorial de palavras

O termo Redes Complexas se refere a um grafo que apresenta uma estrutura topográfica não trivial, composto por um conjunto de vértices (nós) que são interligados por meio de arestas (Barabási, 2003). A teoria das Redes Complexas está relacionada com a modelagem de redes reais, através da análise de dados empíricos. Redes Complexas não são estáticas (evoluem no tempo alterando sua estrutura), e constituem estruturas onde processos dinâmicos (como disseminação de virus ou opiniões) podem ser simulados.

2 Objetivo

Esse trabalho tem como principal objetivo caracterizar a dinâmica de disseminação das notícias no país através de redes complexas.

3 Metodologia

3.1 Rede de Disseminação Real

Nesta primeira parte do trabalho, iremos construir uma rede complexa que descreva o processo epidemiológico de uma notícia. O objetivo da construção dessa rede, é conseguir aproximar a rede de interligação real entre os artigos selecionados do banco MediaCloud, de forma a torná-la observável.

Primeiramente, selecionamos todos os artigos do banco referentes a uma determinada notícia, criando assim, nosso corpus linguístico. Depois ,a partir desse corpus, utlizamos métodos de PLN para representar matricialmente nosso conjunto de dados. A partir da matriz criada construímos nossa rede direcionada de disseminação, onde cada vértice ou nó da rede é um artigo.

Para construir a rede de disseminação precisamos definir a relação de contágio entre artigos, ou seja, o que conecta um artigo ao outro na rede. No nosso caso, definimos que o contágio se dá através de uma relação de influência, porém como definir essa relação? Partindo do pressuposto de que se um artigo influênciou outro então esses artigos são necessariamente similares, consideramos a **similaridade** entre os artigos o fator principal para definir se eles possuem uma conexão em nossa rede.

Possuindo uma conexão, ainda precisamos saber quem influênciou e quem foi o influenciado. Para definir isso, utilizamos o **tempo**. Assim, dado que dois artigos são similares, definimos como influenciador o artigo que fui publicado primeiro e influenciado o artigo publicado depois. No caso de um artigo ter mais de um similar publicado anteriormente a ele, definimos como influenciador o artigo mais similar à ele. Dessa forma todos os vértices de nossa rede possuem um grau de entrada igual à 1 e podem possuir um grau de saída maior do que 1.

A seguir, podemos ver um exemplo do que esperamos da nossa rede de disseminação real:

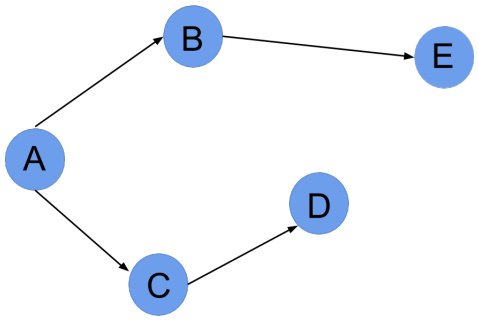

Figura 1: Rede Real Genérica

No exemplo acima, temos que o artigo *A* é similar aos artigos *B* e *C* e possui uma data de publicação mais antiga do que os dois. Sendo assim, foi definido como o influenciador dos dois, ou seja, foi o artigo que os contagiou com a notícia.

3.1.1 Matriz de Documentos

Representamos vetorialmente nosso conjunto de dados utilizando o Word2Vec, uma ferramenta que promove uma implementação eficiente do modelo skip-gram e bag-of-

words contínuo (inserir referencia).

Para gerar nosso modelo word2vec, fornecemos como entrada o corpus linguístico referente a toda a coleção de documentos e o modelo nos retorna uma matriz, onde as linhas são as palavras contidas no corpus e as colunas são os atributos gerados pelo modelo:

(AJEITAR MATRIZ)

aqui talvez seja melhor usar o ambiente array, ao invés de tabular

	a_1	a_2	...	a_m
v_1				
v_2				
...				
v_n				

Temos nessa matriz a representação de todas as palavras presentes em nosso corpus, porém, queremos representar documentos e não palavras.Para isso, buscamos o vetor referente de cada palavra contida em um documento e somamo-os. Dessa forma, associamos todas as palavras contidas em um documento, e transformamo-as em um único vetor. Ou seja, dado um documento A, sabemos que ele é composto pelo seguinte conjunto de palavras $P : \{1,2,3,4,5\}$. Para cada termo buscamos o seu vetor representativo v_t, $t = \{1,2,3,4,5\}$ e somamos todos esses vetores, criando o vetor v_D que representa o vetor do documento D:

como é esta representação?, porque ela é adequada para o seu trabalho? explicar isso em uma seçãoo sobre o skipgram no referencial teórico

$$v_D = v_1 + v_2 + v_3 + v_4 + v_5$$

Representando um documento dessa forma, não levamos em consideração a importância de cada palavra para o documento. Para melhorar nossa representação, antes de somar os vetores, iremos multiplicar cada um pelo valor **Tf-Idf** da palavra a qual ele representa.

O modelo Tf-Idf (*term frequency-inverse document frequency*), é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a um corpus linguístico muito usada para rankeamento de documentos em uma consulta. O Tf-Idf trata-se do produto entre as estatísticas $Tf_{d,t}$ e Idf_t.

mover a explicação do TF-IDF para o a seção 1.2.3, referencie a equação do tf-IDF aqui e deixe na metodologia apenas o que você fez, usando o TF-IDF e outras coisas mais.

Dado um conjunto de N documentos, $Tf_{d,t}$ é a frequência do termo t no documento d, ou seja, o número de vezes em que t ocorre em d. Usamos o termo Tf para computar escores de correspondência consulta-documento, porém, o Tf nos dá a frequência absoluta dos termos, o que faz com que um termo que possua $Tf = 10$ seja 10 vezes mais relevante do um que possua $Tf = 1$.

Podemos concordar que um documento com $Tf = 10$ é mais relevante do que um com $Tf = 1$, porém não necessariamente 10 vezes mais relevante. A relevância não aumenta em proporção com a frequência do termo. Para contornar isso, é comum usar ao invés da frequência absoluta uma ponderação pelo *Log* da frequência. Dessa forma, o peso log da frequência do termo t em d é definido como:

$$W_{t,d} = \begin{cases} 1 + logTf_{t,d} & \text{se } Tf_{t,d} > 0 \\ 0 & \text{caso contrário} \end{cases} \quad (1)$$

Exemplificamos abaixo a correspondência de valores $Tf_{t,d}$ absoluto com a ponderação $W_{t,d}$:

$Tf_{t,d}$	$W_{t,d}$
0	0
1	1
2	1.3
10	2
1000	4

Sabemos que nem todo termo frequente em um documento pode ser considerado muito relevante. Consideramos uma consulta com dois termos: um frequente no conjunto de documentos e outro raro. Não queremos que um documento que possua o termo frequente seja mais relevante do que o documento que possua o termo raro.

Inferimos então que termos raros são mais informativos do que termos frequentes. Dessa forma, queremos dar uma maior relevância para termos raros do que para termos muito frequentes. Para incluir isso em nossa medida usamos o termo Idf.

O termo Idf_t é uma medida de informatividade do termo t, que afeta o rankeamento de documentos para consultas com pelo menos dois termos. Com ele aumentamos o peso relativo de termos raros e diminuimos o peso relativo de termos muito frequentes. O definimos da seguinte maneira:

$$Idf_t = log \frac{N}{df_t}$$

Onde df_t é a frequência de documento, o número de documentos em que t ocorre. Consideramos df_t uma medida inversa da informatividade do termo t.

Ao multiplicarmos o termo Idf ao nosso peso ponderado $W_{t,d}$, temos a medida Tf-Idf. O peso Tf-Idf aumenta com o número de ocorrências dentro de um documento e com a raridade do termo na coleção. É considerado o melhor esquema de ponderação em recuperação da informação.

$$W_{t,d} = (1 + logTf_{t,d}) \cdot log\frac{N}{Df_t}$$

Utilizando essa medida em nosso modelo, temos para cada palavra i um valor $w_{t,d}$, referente ao valor Tf-Idf do termo t no documento d. Sendo assim:

$$v_D = \sum_{i=1}^{5} v_i \cdot w_{t,d}$$

Você copiou este texto de algum lugar?

3.1.2 Construção da Rede

Construimos a rede de disseminação a partir da matriz de documentos gerada na sessão anterior, onde cada vetor consiste em um artigo.

Ao construir a rede dos nossos documentos, consideramos como contágio sofrer influência de outro artigo, ou seja, um artigo A contamina um artigo B se o artigo A influenciou o artigo B.

Dessa forma, em nossa rede, os vértices representam os artigos, e as arestas a relação de influência entre eles, isso é, dado um nó e_i e um nó e_j existe uma aresta a_{ij} que sai de i e vai para j se o artigo i influênciou o artigo j.

Para definir quando um artigo influencia outro em nossa rede, foram usadas duas heurísticas: **similaridade de cosseno** e **temporalidade**.

Utilizamos a similaridade de cossenos para definir se dois artigos são similares ou não. Para isso, calculamos a distância de cossenos entre o ângulo que os dois vetores formam. Podemos calculá-lo utilizando a fórmula abaixo:

$$cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|}$$

Se a distância calculada for suficientemente pequena, consideramos os artigos similares.

Possuindo artigos similares, inferimos que existe uma relação de influência entre eles, porém não temos como saber qual influênciou e qual foi influenciado. Para definir essa questão, utilizamos a data e hora em que cada um foi publicado. Dessa forma, se dois artigos são similares, consideramos como influenciador aquele que foi publicado primeiro.

INCLUIR O QUE É SUFICIENTEMENTE PEQUENO

3.2 Simulação Rede de Disseminação

Nesta parte do trabalho, temos como objetivo simular a disseminação de nossa rede original utilizando o modelo (adicionar referencia) para tentar validar nossa rede de disseminação como um processo epidemiológico.

3.2.1 Rede Completa

Para simular nossa rede, construimos primeiramente uma rede completa com os mesmos nós de nossa rede original, onde os pesos das arestas são as probabilidades dessa aresta existir no caminho de disseminação da notícia, ou seja, a probabilidade de um artigo influenciar um outro. Para isso, levamos em conta o veículo que publicou o artigo. Dessa forma podemos saber através dos veículos qual a chance de um artigo do veículo x ser influenciado por um artigo do veículo y.

Construimos então, uma matriz de pesos identificando todos os veículos presentes na rede original e contando quantas vezes cada veículo x foi influenciado pelo veículo y. Dado que na rede original possuimos n veículos diferentes, temos uma matriz quadrada *nxn*, onde cada posição a_{ij} nos da a quantidade de vezes que o veículo i foi influenciado pelo veículo j. Consideramos que um veículo não influência a si mesmo, logo a diagonal da nossa matriz é de zeros:

$$\begin{array}{c} \\ \text{a} \\ \text{b} \\ \vdots \\ \text{c} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a} & \text{b} & \dots & \text{c} \end{array} \\ \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{31} & a_{32} & \dots & 0 \end{bmatrix}_{nxn} \end{array}$$

Por exemplo, a_{12} é o número de vezes que artigos do veículo *b* influenciaram artigos do veículo *a*.

A partir da matriz de pesos damos os pesos de cada aresta de nossa rede completa. Para cada nó da rede, identificamos seu veículo e atribuimos a todas as arestas que saem dele as chances dele influenciar cada vizinho, e a todas as arestas que chegam nele as chances dele ser influenciado por cada vizinho:

[DESENHO REDE COMPLETA GENERICA]

3.2.2 Simulação

Na simulação da disseminação, utilizamos o modelo epidemiológico [incluir referencia]:

$$\begin{cases} \frac{d\rho_i^I}{dt} = -\rho_i^I(t) + \lambda\rho_i^S(t)\sum_{j=1}^{N} a_{ij}\rho_j^I(t) \\ \frac{d\rho_i^S}{dt} = -\lambda\rho_i^S(t)\sum_{j=1}^{N} a_{ij}\rho_j^I(t) \end{cases} \tag{2}$$

[EXPLICAR MODELO]

Ao terminar a simulação obtemos uma matriz de estados, onde cada posição a_{ij} se refere a probabilidade do artigo j estar contaminado no passo i da simulação. Utilizando uma distribuição de bernoulli, definimos os artigos que estão infectados em casa passo, ficando assim com uma matriz booleana de estados.

A partir da matriz booleana, podemos criar todos os nós de nossa rede simulada. Porém, ainda precisamos criar as arestas, ou seja, definir as relações de influência.

Para definir as relações de influência, a cada novo artigo contaminado no passo i, consideramos como possíveis influenciadores todos os artigos infectados no passo $i-1$. Utilizando a matriz de pesos entre os domínios, calculamos a probabilidade de cada artigo infectado no tempo $i-1$ ter infectado cada novo artigo infectado no tempo i. Definimos o artigo influenciador a partir da probabilidade que ele tem de influenciar o novo artigo infectado.

4 Resultados

4.1 Rede de Disseminação Original

Abaixo temos uma visualização da rede criada:

4.2 Simulação Rede de Disseminação

5 Conclusão

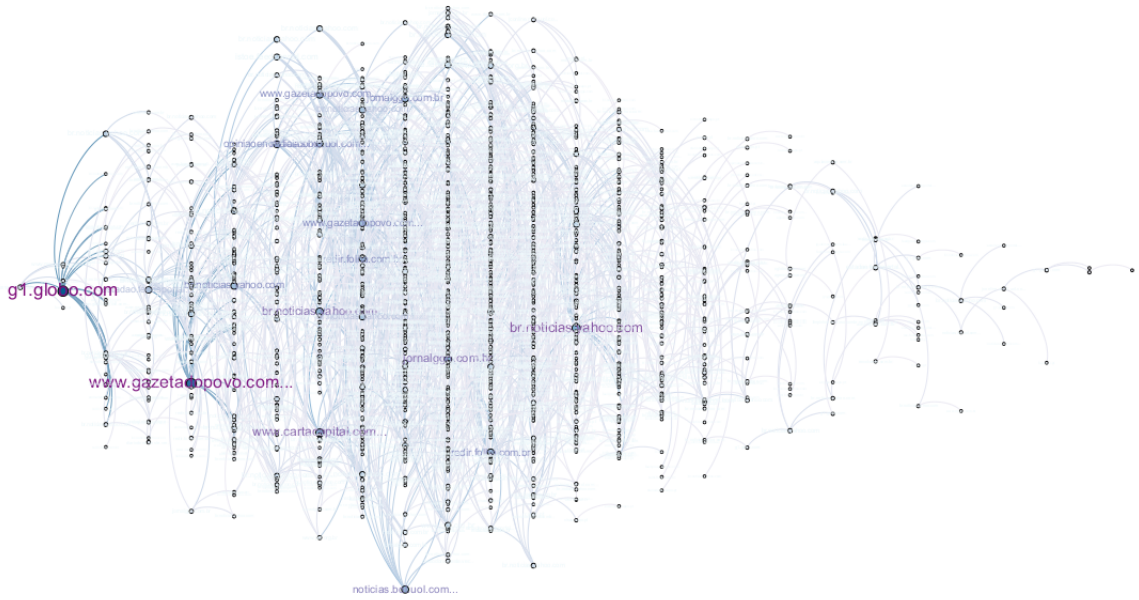

Figura 2: Rede Disseminação