UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

BACHARELADO EM ENGENHARIA DE SOFTWARE

PRISCILA ROCHA FERREIRA RODRIGUES

DINÂMICA DE TEMAS ABORDADOS NO TWITTER VIA EVOLUÇÃO DE CLUSTERS

QUIXADÁ

2016

PRISCILA ROCHA FERREIRA RODRIGUES

DINÂMICA DE TEMAS ABORDADOS NO TWITTER VIA EVOLUÇÃO DE CLUSTERS

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Engenharia de Software da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharel.

Área de concentração: Computação

Orientadora Profa. Ticiana Linhares Coelho da Silva

QUIXADÁ
2016

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Campus de Quixadá

R615d Rodrigues, Priscila Rocha Ferreira
Dinâmica de temas abordados no Twitter via evolução de clusters/ Priscila Rocha Ferreira Rodrigues. – 2016.
58 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Bacharelado em Engenharia de Software, Quixadá, 2016.
Orientação: Profa. Msc. Ticiana Linhares Coelho da Silva
Área de concentração: Computação

1. Cluster (Sistemas de computador) 2. Análise por agrupamento 3. Twitter (Rede social on-line) I. Título.

CDD 005.14

PRISCILA ROCHA FERREIRA RODRIGUES

Dinâmica de Temas Abordados no Twitter via Evolução de Clusters

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Engenharia de Software da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharel.

Área de concentração: computação

Aprovado em: _____ / Janeiro / 2016.

BANCA EXAMINADORA

_____
Profa. MSc. Ticiana Linhares Coelho da Silva (Orientadora)
Universidade Federal do Ceará-UFC

_____
Prof. MSc. Régis Pires Magalhães
Universidade Federal do Ceará-UFC

_____
Prof. Dr. Flávio Rubens de Carvalho Sousa
Universidade Federal do Ceará-UFC

À Deus, por ter feito meu sonho real e ter colocado no meu caminho pessoas que foram fundamentais para o êxito nessa trajetória.

AGRADECIMENTOS

Agradeço a Deus, por ter direcionado meus passos e ter me possibilitado está firme durante todas as etapas da minha vida, sobretudo durante a graduação, onde tantos desafios e responsabilidades surgem ao longo do caminho. Obrigada pela companhia, pela força e discernimento a mim concedidos.

Aos meus pais, Glaide Rodrigues e Geovani Rodrigues, por todo o esforço, toda paciência, todas as orações e toda dedicação que me possibilitaram ter a oportunidade de concluir essa graduação. A todos os meus familiares que torceram para que eu concluísse essa jornada. Vocês são minha grande fonte de motivação e inspiração.

Ao meu namorado, Áquila Alves de Castro, que ao longo de toda a graduação foi meu melhor amigo, e me ofereceu o ombro, me ajudando, apoiando e incentivando em todos os momentos.

Aos professores da Universidade Federal do Ceará campus Quixadá, em especial a minha orientadora, professora Msc. Ticiana Linhares, que para muito além de meramente orientar um trabalho acadêmico, demonstrou também dedicação, competência, amizade e paciência ao longo da produção dessa monografia. Ao professor Msc. Samy Soares, meu tutor no PET, que contribuiu grandemente para minha formação acadêmica e aos professores Drs. Régis Pires, Flávio Rubens e Tânia Pinheiro, que também contribuíram de forma relevante para este trabalho.

Agradeço aos colegas do grupo PET e aos amigos de graduação que sempre estiveram e continuarão presentes, em especial, André Martins, Anderson Uchôa, Beatriz Brito, Evelyne Avelino, Fábio Janyo, Fernanda Amâncio e Otávio Augusto, pelas muitas madrugadas de estudos, brincadeiras e companheirismo ao longo desse curso.

Todos aqui citados me inspiram, me ajudam, me desafiam e me encorajam a ser cada dia melhor. Com vocês, as pausas entre um parágrafo e outro de produção fizeram melhor tudo o que tenho produzido na vida.

“Alguns homens veem as coisas como são, e dizem ‘Por quê? ’ Eu sonho com as coisas que nunca foram e digo ‘Por que não? '” (George Bernard Shaw)

RESUMO

O uso progressivo das redes sociais nos últimos anos tem produzido um grande volume de informações geradas pelos seus usuários, que com frequência compartilham seus sentimentos, opiniões sobre grandes eventos, catástrofes, epidemias, lançamentos de produtos, dentre outros acontecimentos. Compreender como determinados assuntos estão evoluindo nas redes sociais e de que maneira tais assuntos estão se relacionando, pode ser um diferencial para organizações que desejam elaborar melhores estratégias de marketing, publicidade dirigida, obter feedbacks de eventos ou algum produto lançado. No entanto, analisar esse grande volume de dados de forma não automatizada consiste em um problema não trivial. No contexto deste trabalho, aplicam-se técnicas de clusterização para identificar conjuntos de assuntos repercutidos no Twitter, e acompanha-se, por meio de técnicas de evolução de *cluster*, a dinâmica e transição desses assuntos, objetivando alcançar uma visão panorâmica e compreender as motivações de tais evoluções. Além disso, aplicam-se duas medidas de similaridade na clusterização dos dados coletados e analisa-se as semelhanças e diferenças nos resultados obtidos por meio de ambas. O algoritmo de evolução de *clusters* empregado neste trabalho detecta as seguintes transições: Aparecimento, Sobrevivência, Desaparecimento, União, Divisão, Expansão e Retração de *clusters.*

Palavras-chave: Clusterização. Transições de *Clusters*. Redes Sociais.

ABSTRACT

The massive use of social networks in recent years has produced a large volume of information generated by its users, who often share their feelings, opinions about major events, disasters, epidemics, product launches, among other events. Understanding how specific subjects are evolving in social networks and how they are related each other, it can be relevant for organizations that aim to make better marketing strategies, advertising targeting, get feedbacks events or any released product. However, to analyze this large volume of non-automated data consists of a non-trivial problem. In the context of this work, we apply a clustering technique to identify sets of issues in Twitter, and we also apply cluster evolution to discover how these issues are dynamic and the transition to each other. We aim to understand and discuss about such cluster evolution. In addition, apply two similarity measures the clustering of data collected and analyzed similarities and differences in the results obtained by both. In this work, we detected the following transitions for clusters: Creation, Survival, Disappearance, Merge, Split, Expansion and Shrinkage of clusters.

Keywords: Clustering. C*lusters* Evolution. Social Networks

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com o crescente uso das redes sociais nos últimos anos, o volume de informações que circulam por meio delas vem aumentando de maneira considerável e tornando tais redes o foco de diferentes estudos. Apenas no Twitter, são mais de 600 milhões de contas ativas e uma média de 58 milhões de *tweets* por dia[1]. Dessa forma, é possível observar uma nova maneira de participação da sociedade, onde os usuários utilizam com frequência as redes sociais para expressar opiniões, sentimentos e compartilhar informação de maneira instantânea com seus seguidores.

Cada vez mais se faz necessário fornecer às organizações e demais usuários uma forma eficaz de extrair conteúdo significativo que permita acompanhar a evolução de eventos ou assuntos repercutidos nas redes sociais. Devido a grande quantidade de conteúdo, o acompanhamento e análise desses dados de forma não automatizada consiste em um problema não trivial. Nesse contexto, a mineração de dados surge como uma grande aliada da descoberta de conhecimento, apresentando um conjunto de técnicas que permite analisar grandes quantidades de dados tentando encontrar um padrão consistente, sumarizar os dados, extrair conhecimento ou realizar predições (SILVA e COELHO DA SILVA, 2014).

Dentre as técnicas que podem ser utilizadas para a mineração de dados, está a clusterização, que consiste no agrupamento de um aglomerado de dados multidimensionais num conjunto de classes, denominadas *clusters*, com base no grau de similaridade das observações (JAIN et al., 1999). Esse grau de similaridade dos dados, fundamental para a construção de um *cluster*, é encontrado por meio das medidas de similaridade adotadas na técnica de clusterização empregada, permitindo a comparação de padrões para o agrupamento de dados em diferentes *clusters*.

Um *cluster* pode ser visto como um conceito que, em virtude da constante evolução e consequentes impactos nos mecanismos de geração dos dados, pode sofrer alterações (KAUR et al., 2009). A constatação da volatilidade dos dados ao longo do tempo, aliada à consciência de que é mais relevante compreender a dinâmica e a natureza da evolução, do que meramente identificá-la (SPILIOPOULOU et al., 2006), foi a principal razão que motivou à escolha do tema em estudo.

No contexto deste trabalho, aplicam-se técnicas de clusterização para identificar conjuntos de assuntos populares repercutidos no Twitter, e acompanhar, por meio de técnicas

---

[1] http://www.statisticbrain.com/twitter-statistics/

de evolução de *cluster*, a dinâmica e transição desses assuntos, objetivando alcançar uma visão panorâmica e compreender as motivações de tais evoluções. Com a aplicação de tais técnicas, é possível detectar não apenas o que está acontecendo em um momento específico, mas principalmente como determinado assunto, evento ou acontecimento está evoluindo e quais os desdobramentos que ele tomou ao longo do tempo.

Tendo em vista a importância das medidas de similaridade para uma melhor exatidão na detecção e evolução de *clusters*, aplicam-se também duas medidas de similaridade nos dados coletados e analisa-se a precisão que ambas apresentam em relação aos acontecimentos no período da coleta, notando semelhanças e diferenças nos resultados obtidos. Foram aplicadas as medidas de similaridade de Jaccard e *Fading*, recentemente empregadas no contexto de redes sociais em YIN et al (2012) e LEE et al (2014) respectivamente.

Dentre outras abordagens, este tipo de estudo pode ser direcionado para a descoberta e observação do modo como os grupos sociais tendem a evoluir por meio de uma dimensão temporal, favorecendo tarefas como a publicidade dirigida e a personalização de conteúdos e serviços, adequando-os às necessidades e preferências dos consumidores. OLIVEIRA e GAMA (2010) citam também a aplicabilidade da abordagem de evolução de *clusters* na Detecção de Fraudes e acompanhamento das tendências na Economia.

Um dos trabalhos base para a proposta deste é o de LEE et al (2014). Nele, foi proposto um *framework* para acompanhar a evolução de *clusters* em bases de dados altamente dinâmicas, tais como as de redes sociais. Um estudo de caso apresentado pelos autores utilizando dados do Twitter possibilitou a detecção de vários eventos e apresentou como estes eventos evoluíram e passaram a se relacionar ao longo dos dias, bem como será realizado neste trabalho. No entanto, diferentemente de LEE et al (2014), no trabalho aqui realizado foram aplicadas técnicas de PLN (Processamento de Linguagem Natural) nos dados coletados, o que possibilitou uma melhor estruturação dos dados e consequentemente uma maior precisão nos resultados. Além disso, como citado anteriormente, foram aplicadas duas medidas de similaridade na clusterização a fim de analisar as semelhanças e diferenças nos resultados apresentado por ambas.

Na próxima seção, são apresentados os trabalhos relacionados. Posteriormente, será exposta a fundamentação teórica, contendo conceitos importantes para a compreensão deste trabalho. Na sequência, é descrito o passo a passo do processo de execução, apresentando os procedimentos metodológicos. Em seguida, serão detalhados os resultados obtidos e, por fim, as considerações finais do trabalho em questão serão abordadas.

2 TRABALHOS RELACIONADOS

Estudos prévios já foram realizados para a compreensão da evolução de *clusters*. A seguir, alguns desses estudos serão apresentados brevemente com o intuito de esclarecer de que maneira se relacionam com o trabalho proposto.

KIM e HAN (2009) propõem um método para acompanhar a evolução de *cluster* em redes de dados dinâmicos, utilizando o conceito de *snapshots* (conjunto de dados em um ponto específico no tempo). No estudo de KIM e HAN (2009), a proposta é decompor uma rede dinâmica em uma série de *snapshots* para momentos distintos no tempo, depois aplicar algoritmos de mineração em cada *snapshot* para encontrar padrões úteis, e por fim, combinar esses padrões entre os diferentes momentos para gerar uma sequência de padrões dinâmicos.

Tendo em vista o potencial em capturar fenômenos naturais e sociais ao longo do tempo, as redes de dados dinâmicos também são o foco de estudo neste trabalho, onde, de maneira semelhante a KIN e HAN (2009) são aplicadas técnicas de clusterização e evolução de *cluster* a fim de detectar padrões e acompanhar a evolução dos *clusters* detectados. Entretanto, a proposta de KIM e HAN (2009) distingue-se do trabalho aqui apresentado pelo fato de que, embora possa lidar com o aparecimento, expansão, retração e desaparecimento de *clusters*, não oferece suporte à atividades como divisão e união de *clusters*, que, segundo LEE et al (2014), são transições bastante comuns em dados dinâmicos e de elevada importância para a completa compreensão das evoluções sofridas pelos *clusters*.

TANG et al (2013), ressalta o grande volume de dados espaço-temporais produzidos por meio de tecnologias móveis e pelo frequente uso de sensores GPS (*Global Positioning System*) e aplicativos como o FourSquare[2]. Tendo isso em vista, os autores propuseram um modelo de clusterização baseado em um padrão de mobilidade conhecido como *traveling companion*. A utilização do padrão de mobilidade acelera o processamento dos dados para criação dos *clusters*, além de também descobrir objetos que viajam juntos. O trabalho utilizou ainda técnicas de evolução de *clusters* para tal finalidade.

A equivalência com este trabalho está no uso de técnicas de clusterização, que foram aplicadas por TANG et al (2013) para identificar grupos de objetos móveis, e também, na aplicação de técnicas de evolução de *cluster*, utilizadas na pesquisa citada para detectar os objetos que viajam juntos. No entanto, uma vez que em seu trabalho TANG et al (2013) direciona suas pesquisas para um cenário de dados de trajetória, neste trabalho as técnicas de

[2] https://pt.foursquare.com/

clusterização e evolução de *cluster* são utilizadas no contexto de redes sociais, que requer a aplicação de procedimentos específicos para o contexto de dados em formato de texto que são escritos em linguagem informal.

Em LEE et al (2014), foi construída uma base de dados direcionada para o domínio de tecnologia, por meio de uma coleta de *tweets* ao longo de um mês. O algoritmo de clusterização DBSCAN foi aplicado aos dados coletados e identificou-se os eventos de Tecnologia que estavam sendo repercutidos no período da coleta, tais como: Sopa Wikipedia Blackout, Apple iBooks, CES Conference e outros. Aplicando posteriormente os algoritmos ICM (Incremental Cluster Maintenance) e eTrack (Cluster Evolution Tracking), propostos pelos autores, foi viável identificar como eventos distintos passaram a relacionar-se ao longo do tempo e encadearam em outros assuntos relacionados, mostrando desta forma a dinâmica dos assuntos abordados no Twitter no período da coleta.

A proposta de LEE et al (2014) assemelha-se a este trabalho por utilizar o Twitter como fonte de extração de dados, além do uso do algoritmo de clusterização DBSCAN, também utilizado neste trabalho para o agrupamento e detecção de assuntos e eventos repercutidos. Outra semelhança está no uso de técnicas de evolução de *clusters*, o que permitiu o acompanhamento da evolução dos eventos detectados na rede social.

No entanto, o trabalho aqui proposto utiliza técnicas de PLN (Processamento de Linguagem Natural) sobre os dados coletados. Tais técnicas não foram empregadas por LEE et al (2014). Como explicado melhor adiante, o uso de PLN é importante por possibilitar um maior refinamento dos dados, removendo inconsistências, palavras irrelevantes e impurezas comuns em dados textuais de escrita livre como os de redes sociais. Outra diferença é que o algoritmo de clusterização DBSCAN é aplicado nos dados coletados para este trabalho de duas maneiras distintas. Uma utilizando a medida de similaridade *Fading*, medida esta utilizada no trabalho de LEE et al (2014), e outra com a medida de Jaccard. As medidas são aplicadas com o intuito de analisar as semelhanças e diferenças que ambas apresentam nos resultados obtidos e verificar quais resultados mais condizem com os acontecimentos reais. É válido ressaltar que no trabalho de LEE et al (2014), a evolução de *clusters* é detectada de forma incremental, o que não ocorre no trabalho em questão. Além disso, as bases de dados aqui utilizada são referentes a assuntos repercutidos no ano de 2015.

A Tabela 1 a seguir, resume as semelhanças e diferenças citadas. É válido destacar que o trabalho aqui apresentado, não abrange todo o contexto englobado nos demais trabalhos. Um exemplo disso é o fato de não ser realizada a manutenção incremental dos clusters, característica abordada em Lee et al 2014. O intuito da Tabela 1 é apenas ressaltar as

características que o trabalho em questão apresenta como diferencial positivo em relação aos demais.

Tabela 1 - Diferencial positivo deste trabalho em comparação com os Trabalhos Relacionados

	Todas as Evoluções	***Contexto: Redes Socias***	***Processamento de Linguagem Natural***	***Compara Medidas de Similaridade***
[Kim and Han, 2009]	NÃO	SIM	SIM	NÃO
[Tang et al, 2013]	SIM	NÃO	NÃO	NÃO
[Lee et al, 2014]	SIM	SIM	NÃO	NÃO
Este Trabalho	**SIM**	**SIM**	**SIM**	**SIM**

3 FUNDAMENTAÇÃO TEÓRICA

A seguir serão descritos importantes conceitos para a compreensão deste trabalho. Na seção 3.1, é apresentada a rede social escolhida como fonte de extração de dados para este trabalho: o Twitter. Em seguida, na seção 3.2, a Mineração de Textos será abordada explanando também a técnica de Processamento de Linguagem Natural. Na seção 3.3 será apresentado o conceito de clusterização bem como os algoritmos que serão utilizados para tal finalidade. Por fim, na seção 3.4 será exposto a definição de Evolução de *Cluster* e explanado de que maneira este conceito será aplicado nos dados coletados do Twitter após a clusterização.

3.1 Twitter

Criado em 2006 por uma *startup* americana conhecida como Obvious (FARHI, 2009), o Twitter desponta hoje como uma das redes sociais mais populares do mundo[3]. É permitido aos usuários do Twitter publicar mensagens de texto curtas, limitadas a 140 caracteres conhecidas como *tweets*. Em função da quantidade reduzida de caracteres, os *tweets* são com frequência escritos em uma linguagem objetiva, informal e abreviada. Rodrigues et al (2012) ressalta que esse modelo de interação rápida do Twitter induz os usuários a se expressar e compartilhar com frequência informações, opiniões e sentimentos.

Segundo RUSSEL (2013), um dos grandes diferenciais do Twitter quanto a outras redes sociais populares seria sua estrutura dinâmica e seu modelo assimétrico de seguidores, onde a relação de seguir e ser seguido não requer reciprocidade. Dessa forma, as conexões entre os usuários são realizadas de maneira unilateral, possibilitando-os amplo acesso as informações publicadas, sem que para isso, tais usuários estejam restritos a possuir alguma permissão de conexão entre eles.

O Twitter disponibiliza uma API (*Application Programming Interface*) REST (*Representational State Transfer*) que permite aos desenvolvedores acessarem os dados dos usuários, atualizações, status e outras informações. REST é um design arquitetural construído para servir aplicações em rede, definindo os recursos e as maneiras de localizar e acessar dados (FIELDING, 2000). Outra API disponibilizada pelo Twitter é a *Streaming* API, responsável por permitir acesso em tempo real ao grande volume de informações disponíveis.

---

[3] http://www.statisticbrain.com/social-networking-statistics/

As aplicações baseadas no Twitter poderão usar tais APIs de forma individual ou combinadas para alcançar os objetivos planejados.

Considerando a grande quantidade de usuários ativos, o alcance mundial desta rede, as API's disponibilizadas e a grande quantidade de informação circulada a cada instante, o Twitter foi escolhido como fonte de exploração de dados para este trabalho. Além das características já citadas, foi também levado em consideração para a escolha desta rede social a quantidade limitada de caracteres, que obriga o usuário a expor o teor do assunto em poucas palavras, facilitando o processo de mineração dos dados por diminuir a quantidade de impurezas e irrelevâncias presentes em textos maiores.

3.2 Mineração de Textos

Vista como uma extensão ou sub-área da Mineração de Dados, a Mineração de Textos é um processo que utiliza algoritmos capazes de analisar coleções de dados textuais com o intuito de extrair conhecimento. Esse processo de Descoberta do Conhecimento por meio da Mineração de Textos, pode ser dividido em etapas que vão desde a seleção dos dados a serem coletados, até a análise e estudo do conhecimento gerado, conforme apresenta a Figura 1.

Figura 1- Etapas do Processo de Mineração de Texto

Fonte: Fonte: Aranha (2007).

ARANHA (2007), explica que na fase da coleta utiliza-se web *crawlers*, programas que visitam sítios e capturam os textos que serão utilizados para a extração de conhecimento.

No pré-processamento, são utilizadas técnicas como o Processamento de Linguagem Natural para estruturar os textos que serão analisados. A indexação é a fase onde são extraídos conceitos dos documentos por meio da análise de seu conteúdo e traduzidos em termos da linguagem de indexação. Na etapa de mineração, são empregadas técnicas e algoritmos para o reconhecimento de padrões ou tendências e extração de conhecimento. Na análise, os resultados são avaliados e validados. Neste trabalho, todas estas etapas são aplicadas, exceto a indexação, pois não foi considerada relevante neste contexto, tendo em vista que não houve necessidade de realizar consultas nos textos que compõem a base de dados.

Segundo HEARST (1999), os dados textuais abrangem uma vasta e rica fonte de informação, embora estejam em um formato em que seja mais difícil de extrair de maneira automatizada, se comparados aos dados organizados em banco de dados. Isso se deve ao fato de que no banco de dados os dados geralmente estão organizados e estruturados em tabelas e as mesmas possuem relações entre si, o que nos garante mais coesão nas informações.

As redes sociais têm gerado uma grande massa de informação textual desestruturada, e neste contexto a Mineração de Texto surge como uma maneira de dar estruturação aos dados textuais, visando facilitar a extração de conhecimento dos respectivos dados.

3.2.1 Processamento de Linguagem Natural

O Processamento de Linguagem Natural é um conjunto de técnicas teórico-computacionais de grande relevância para Mineração de Textos. Empregando conhecimentos da área de linguística, o PLN possibilita explorar ao máximo o conteúdo do texto, extraindo entidades, seus relacionamentos, identificando sinônimos, corrigindo palavras escritas de forma errada e ainda desambiguizando-as.

LIDDY (2001) mostra que para processar a linguagem natural, o PLN a representa em diversos níveis, como léxico, morfológico, semântico, etc. LIDDY (2001) ressalta ainda a complexidade deste processo por lidar com diversos elementos linguísticos e estruturas gramaticais.

No contexto deste trabalho, foi aplicada a remoção de *stopwords*, uma técnica de Processamento de Linguagem Natural para remoção de palavras dotadas de pouco valor semântico e com elevada recorrência em qualquer texto. São os pronomes, artigos, advérbios, e outros termos considerados irrelevantes por não serem palavras que expressam conteúdo significativo dentro do *tweet*.

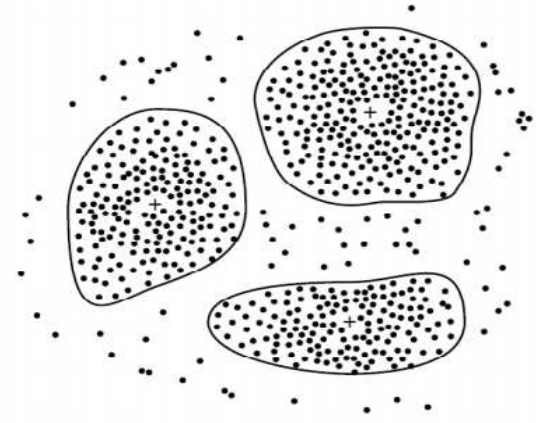
A utilização desta técnica proporcionou maior consistência e estruturação aos *tweets* que foram analisados, tendo em vista que as palavras que não possuem relevância para os

resultados da clusterização e evolução dos *clusters* foram descartadas dos conjuntos de dados, melhorando o desempenho e precisão do algoritmo de clusterização utilizado.

3.3 Clusterização

Com base em suas inúmeras pesquisas na área de agrupamento de dados, JAIN (2010) define clusterização como o estudo formal dos métodos e algoritmos para agrupamento de objetos com base nas semelhanças ou nas características intrínsecas percebidas. Conforme apresenta a Figura 2, a tarefa de agrupamento visa identificar e aproximar os registros similares em grupos disjuntos chamados *clusters*.

Figura 2-Dados agrupados em 3 *clusters*



Fonte: Han et al (2006).

Os algoritmos de clusterização são classificados de acordo com as diferentes técnicas que empregam no agrupamento de dados. Neste trabalho foram utilizados métodos de clusterização por densidade, onde os *clusters* são tratados como regiões de elevada densidade, sendo separados de regiões de baixa densidade. A possibilidade de identificar *clusters* de maneira arbitrária e o fato de não necessitar da definição do número de *clusters* como parâmetro inicial (YIP et al., 2006) são as principais vantagens dessa abordagem. Nesse contexto, o algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) proposto em ESTER et. al (1996) é um dos mais referenciados da literatura. Trata-se de um algoritmo simples, eficiente, e que contempla conceitos importantes, que servem de base para qualquer abordagem baseada em densidade, sendo por isso escolhido para aplicação neste trabalho.

Este algoritmo utiliza uma abordagem baseada em centro, onde cada dado de um *cluster* tem uma vizinhança de um determinado raio, que contém um número mínimo de

pontos. O DBSCAN recebe como entrada um conjunto de dados *X*, o tamanho da vizinhança denominado *eps* e o número mínimo de pontos de uma vizinhança, denotado *minPts*. Classifica os dados em *core*, *border* ou *outlier*. Um ponto é *core* se possuir uma vizinhança maior ou igual do que o estabelecido em *minPts*. É considerado *border* se pertence à vizinhança de um ponto *core*, ou seja, se a distância entre eles é menor que *eps*. Pontos *outliers* são pontos atípicos ou inconsistentes que, por não pertencerem a *cluster* nenhum, são descartados (ESTER et al., 1996).

Neste trabalho a clusterização foi aplicada nos dados coletados do Twitter com o intuito de identificar *clusters* de assuntos repercutidos na rede social durante o período da coleta de dados.

3.3.1 Medidas de Similaridade

Chamam-se Medidas de Similaridade as métricas que permitem comparar padrões pertencentes a um espaço de características, sendo tais medidas, segundo KASZNAR e GONÇALVES (2009), as responsáveis por grande parte dos esforços empregados no processo de clusterização. Se dois padrões são similares de acordo com a medida utilizada pela técnica de clusterização empregada, então serão agrupados em um mesmo *cluster*, caso contrário, serão agrupados em *clusters* distintos.

Nesta subseção serão abordadas duas medidas de similaridade: a similaridade de Jaccard e a medida de similaridade *Fading*. Tais medidas foram respectivamente aplicadas nos recentes trabalhos de YIN et al (2012) e LEE et al (2014), ambas no contexto de mineração de dados em redes sociais.

A medida de similaridade de Jaccard consiste no coeficiente obtido por meio do tamanho da intersecção pelo tamanho da união entre conjuntos de dados, definido como:

$$J\left(p^L, b^L\right) = \frac{\left| p^L \cap b^L \right|}{\left| p^L \cup b^L \right|}$$

Quanto mais próximo de 1(um) esse valor se encontra, mais semelhantes são esses conjuntos de dados. Desta forma, considerando o conjunto de dados A com 5 objetos e o conjunto de dados B com 6 objetos, se 3 elementos entre eles são similares sofrendo assim intersecção e sabendo que a união entre esses conjuntos é consequentemente 8, logo a similaridade de J(A,B) = 3/8.

Considerando o contexto de redes sociais, LEE et al (2014) pondera que a medida de tempo se faz importante tendo em vista que publicações mais próximas no tempo tem maiores possibilidades de tratar sobre o mesmo assunto. Assim, propõe a medida de similaridade *Fading* para capturar tanto semelhanças no conteúdo das postagens como semelhanças no tempo. Formalizando p_i^L como uma lista de palavras em um momento i, p_j^L como uma lista de palavras em um momento j no tempo, LEE et al (2014) usa uma função exponencial para incorporar o efeito do tempo decorrido entre as publicações, sendo a similaridade *Fading* definida como:

$$S_F(p_i, p_j) = \frac{|p_i^L \cap p_j^L|}{|p_i^L \cup p_j^L| \cdot e^{|p_i^\tau - p_j^\tau|}}$$

As duas medidas de similaridade foram aplicadas nos dados estudados neste trabalho de maneira dissociada, com o intuito de que os resultados apresentados por ambas fossem validados e comparados.

3.4 Evolução de *Clusters*

Devido a natureza dinâmica da maioria dos conjuntos de dados, emergiram novos métodos e técnicas para manter e atualizar conhecimento anteriormente descoberto. Essa incessante evolução dos dados exige a adoção de novas perspectivas orientadas para o tempo. O tempo surge assim, como uma dimensão adicional através do qual os dados podem evoluir e sofrer mudanças (OLIVEIRA et al 2010).

A metodologia MONIC (*Modeling and Monitoring Cluster Transitions*), proposta por SPILIOPOULOU et al. (2006), apresenta uma clara tipificação das transições que os *clusters* podem experimentar. Analisando o *cluster* como membro integrante de um esquema mais geral que é a Clusterização, os autores preveem os seguintes tipos de transições entre eles: Aparecimento, Sobrevivência, União, Divisão, Expansão, Retração ou Desaparecimento. Ou seja, o *cluster* pode não sofrer nenhuma alteração e simplesmente sobreviver, pode crescer ao longo do tempo e sofrer uma expansão, pode se unir a outro *cluster* ao apresentar características similares, pode dividir-se em mais de um *cluster* ou pode simplesmente desaparecer. A Tabela 2 apresentada abaixo, sintetiza as transições propostas por

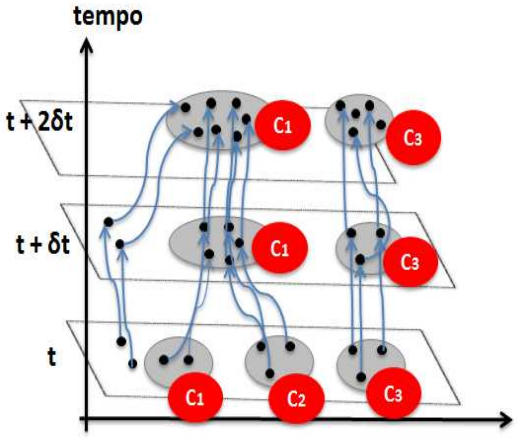
SPILIOPOULOU et al. (2006). Esta é uma abordagem muito próxima da que será utilizada neste trabalho.

Tabela 2 - Evoluções de um *cluster*

Transição	***Notação***	***Definição***
Sobrevivência	$S \rightarrow S'$	S' manteve todos os elementos de S
Aparecimento	$\odot \rightarrow S$	Nenhum *cluster* é similar a S no *timestamp* anterior
Desaparecimento	$S \rightarrow \odot$	Nenhum *cluster* é similar a S no *timestamp* seguinte
Divisão	$S \rightarrow \{S'1, \ldots, S'j\}$	O *cluster* S é similar a um conjunto de *clusters*
União	$\{S1, \ldots, Sj\} \rightarrow S'$	Um conjunto de *clusters* é similar ao *cluster* S'
Expansão	$S \nearrow S'$	S' é similar a S, porém com quantidade de elementos maior que S
Retração	$S \searrow S'$	S' é similar a S, porém com quantidade de elementos menor que S

Imagine o seguinte cenário, ilustrado na Figura 3. No tempo t, existem três *clusters* e alguns *outliers* no conjunto de dados. Considere que cada ponto é uma postagem (*tweet*) e cada *cluster* representa uma região densa com postagens sobre o mesmo assunto, por exemplo. Após δt unidades de tempo, as postagens do *cluster* C1 e C2 começam a tratar sobre o mesmo assunto, e dessa forma, se unem. Observe que na Figura, foram renomeados todos os objetos do *cluster* C2 como *cluster* C1. Isso acontece em cenários reais quando, por exemplo, assuntos distintos acabam se relacionando ao longo do tempo. Note que no *cluster* C3, mais objetos foram adicionados ao conjunto de dados entre os tempos t+ δt e t+ 2δt. O *cluster* C3 no tempo t+ 2δt pode representar, por exemplo, o aumento do número de *tweets* sobre um evento, representando assim a expansão da popularidade e repercussão deste evento. Ambos os fenômenos apresentados caracterizam-se como evolução de *clusters*.

Figura 3 - Registro de dados em 3 momentos no tempo



Fonte: Coelho da Silva et al (2014)

As técnicas de evolução de *clusters* foram aplicadas neste trabalho com o intuito de detectar as transições que os assuntos têm sofrido ao longo do tempo no Twitter e de que maneira estão repercutindo e se relacionando com outros temas.

4 OBJETIVOS

4.1 Objetivo Geral

Tendo em vista a dinamicidade dos assuntos abordados nas redes sociais, o objetivo deste trabalho consiste em apresentar, de maneira automatizada, a evolução de acontecimentos repercutidos no Twitter ao longo de um determinado período de tempo, compreendendo e identificando as correlações entre eles.

4.2 Objetivos Específicos

- Selecionar assuntos relevantes ao longo do primeiro semestre de 2015 e coletar esses dados textuais (*tweets*) do Twitter.
- Realizar o pré-processamento dos dados coletados, removendo inconsistências e impurezas dos mesmos por meio do Processamento de Linguagem Natural.
- Aplicar a técnica de clusterização nos dados coletados a fim de identificar grupos de assuntos distintos.
- Acompanhar a dinâmica dos assuntos, apontando de que maneira tais assuntos têm repercutido e como se modificaram ou se combinaram a outros temas.
- Gerar nuvens de palavras e gráficos que forneçam uma visão panorâmica das evoluções sofridas pelos assuntos ao longo do tempo.
- Comparar o resultado obtido entre as duas medidas de similaridade adotadas na clusterização;
- Interpretar e validar o resultado obtido.

5 PROCEDIMENTOS METODOLÓGICOS

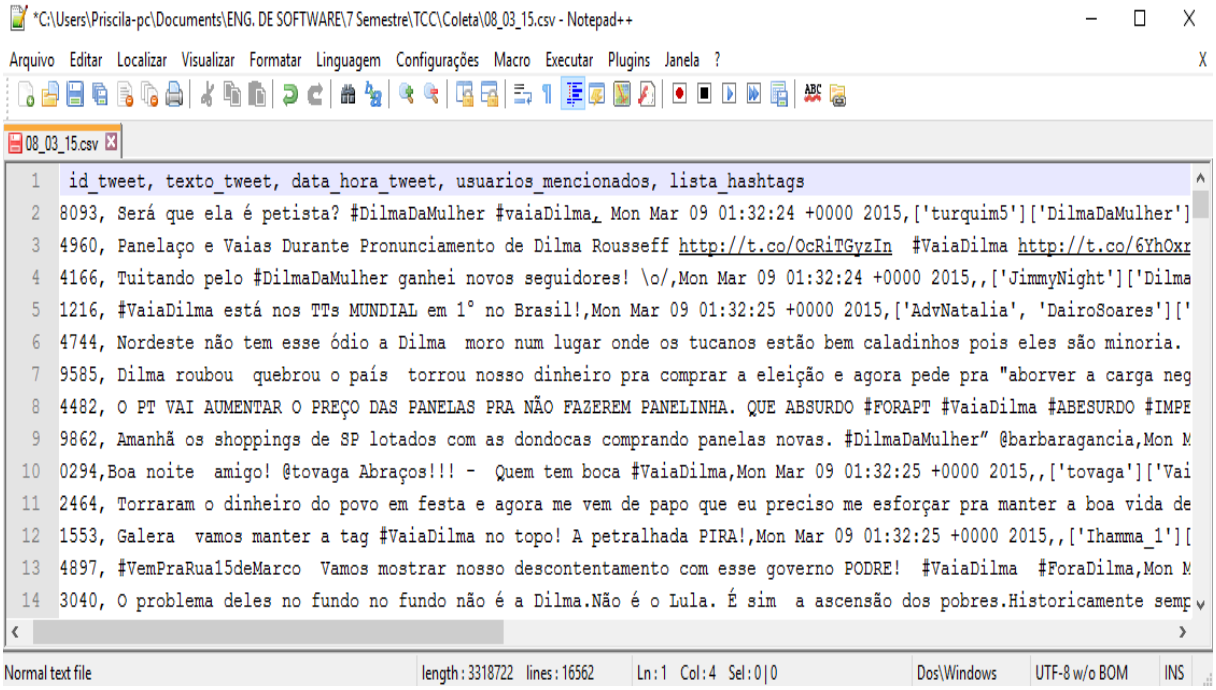
5.1 Coleta de *tweets*

A primeira etapa da execução deste trabalho consiste na coleta dos dados do Twitter. Para esta fase, foi implementado um script escrito na linguagem Python. O *script* recebe como parâmetro palavras e retorna *tweets* que as possuem. Esses *tweets* são em seguida armazenados em um arquivo CSV (*Comma-separated values*), arquivo que contém dados sequenciais em cada linha, separados por um caractere de separação (em geral, uma vírgula ou um ponto e vírgula), conforme pode ser visto na Figura 4.

No contexto deste trabalho, foi selecionada uma palavra para cada assunto que se pretendia coletar. Para a coleta de dados relacionada aos protestos existentes no país pedindo o *impeachment* da presidente, a palavra "Dilma" foi escolhida para passar como parâmetro para o algoritmo, por exemplo.

Cada *tweet* possui uma série de atributos, como id, texto, número de vezes que foi curtido, coordenadas, dentre outros. Para este trabalho, foram armazenados no arquivo CSV apenas os atributos considerados relevantes para a execução do mesmo. São eles: id do *tweet*, texto do *tweet*, lista de *hashtags* (palavras-chave precedidas do símbolo #), data de criação e horário do *tweet* e os usuários mencionados no *tweet.* O texto, o horário e a data do *tweet* são atributos essenciais para a etapa de análise da evolução dos *clusters.*

Figura 4-CSV gerado na coleta de 08/03/2015



Fonte: Elaborado pelo Autor.

Este processo de coleta ocorreu entre março e junho de 2015, posteriormente iniciou a etapa de pré-processamento, explicada na subseção a seguir.

5.2 Pré- processamento dos *tweets* coletados

Nesta etapa, os dados coletados passam por um pré-processamento e os textos são estruturados, para que assim possam ser minerados. Nesta etapa é aplicada a técnica de remoção de *stopwords*, onde palavras dotadas de pouco valor semântico e que não agregam valor à análise dos dados foram removidas. São artigos, preposições, conjunções, dentre outros, como: as, os, em, de, para, com, foi, e etc. Foram descartados também links e caracteres não alfabéticos, ficando desta forma, apenas as palavras chaves do *tweet* e que de fato representem conteúdo significativo para as etapas de clusterização e evolução de *clusters*.

Para aplicação da técnica de remoção de *stopwords* foi utilizada uma biblioteca chamada *Natural Language Toolkit* (NLTK). NLTK é uma biblioteca para a linguagem Python distribuída gratuitamente sob a licença open source com a finalidade de pesquisa e desenvolvimento na área de Processamento de Linguagem Natural (PLN). Este kit possui uma extensa documentação, tanto on-line quanto em formato pdf, disponibilizada gratuitamente no site dos desenvolvedores do NLTK[4].

Depois de realizada esta estruturação nos dados coletados, seguimos para a clusterização, utilizando o algoritmo DBSCAN. Esta etapa é detalhada na próxima subseção.

5.3 Clusterização dos dados

Nesse ponto, foi aplicada a técnica de clusterização aos dados coletados do Twitter. Para isso, o algoritmo de clusterização DBSCAN foi implementado e aplicado. A implementação do DBSCAN ocorreu de duas maneiras distintas: uma utilizando a medida de similaridade de Jaccard, e outra utilizando a medida de similaridade *Fading*, explicadas anteriormente na seção 2.2.1. Ambas as implementações foram aplicadas em uma série de *snapshots* para distintos momentos no tempo, a fim de encontrar padrões e detectar grupos de assuntos que surgem ou permanecem ao longo do tempo.

Aplicada a clusterização, os dados estavam prontos para o emprego da técnica de evolução de *clusters*, explicada na subseção a seguir.

---

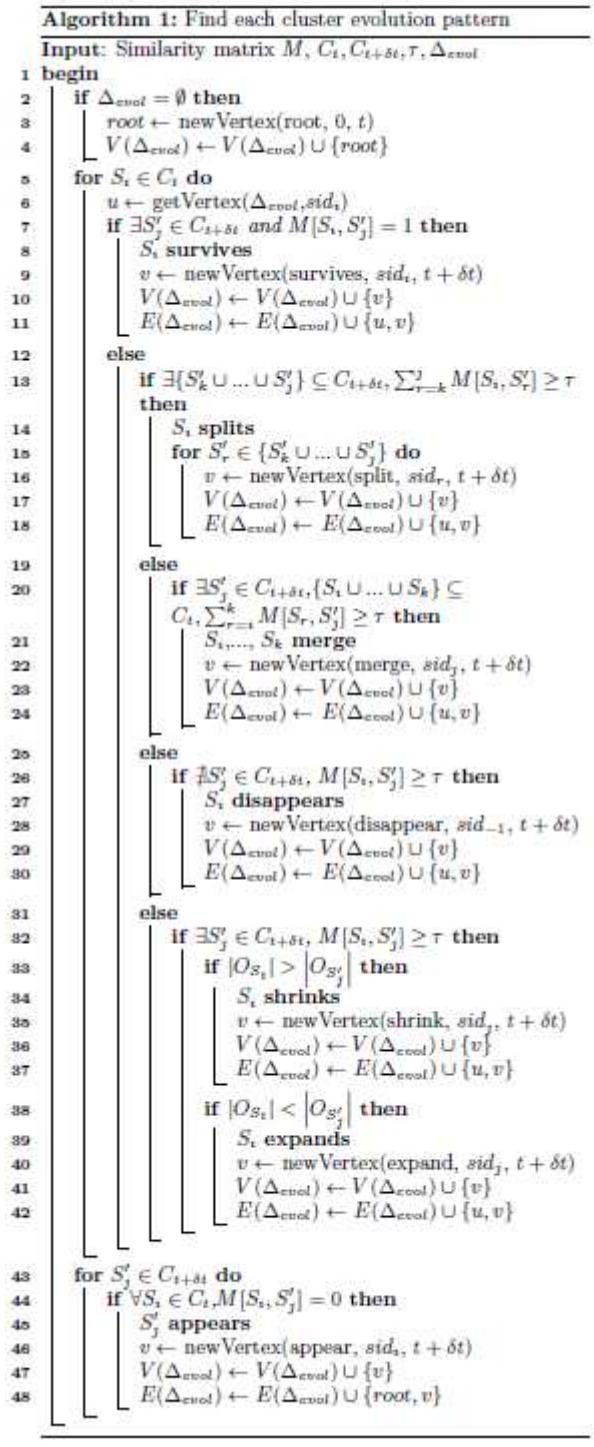
[4] http://nltk.org/

5.4 Evolução de *Clusters*

Para realizar o monitoramento da evolução de *clusters*, foi executada uma implementação do Algoritmo 1 apresentado a seguir. Proposto por COELHO DA SILVA et al (2014), o Algoritmo 1 recebe como entrada uma matriz M onde foram armazenados em cada célula o valor de similaridade entre dois *clusters* em diferentes momentos do tempo. O Algoritmo 1 recebe também como entrada C_t, que são todos os *clusters* encontrados no tempo t, e $C_{t+\delta t}$, todos os *clusters* encontrados no tempo t + δt. Os outros dois parâmetros passados são o τ e o $\Delta_{evol.}$ O τ consiste no *threshold* de similidaridade, ou seja, o limiar que define o número mínimo de elementos em comum que dois conjuntos devem possuir para serem considerados similares. O Δ_{evol} é o grafo de evolução gerado de forma incremental pelo algoritmo.

Após a primeira estrutura de repetição, o Algoritmo 1 se divide em 6 blocos no qual o primeiro é capaz de detectar quando um *cluster* sobrevive entre um momento e outro, nesse caso a similaridade entre o *cluster* S e S' é 1 (ou seja, S' manteve todos os elementos de S). O segundo bloco identifica quando ocorre divisão em um *cluster*, nesse caso o *cluster* S é similar a um conjunto de *clusters* {S'1,...,S'j} obtidos no *timestamp* seguinte ao de S. O terceiro detecta uma união entre *clusters*, nesse caso um conjunto de *clusters* {S1,...,Sj} é similar ao *cluster* S' obtido no *timestamp* seguinte. O quarto bloco detecta o desaparecimento de um *cluster*, nesse caso nenhum *cluster* é similar a S no *timestamp* seguinte. O quinto e sexto bloco detectam, respectivamente, quando um *cluster* diminui ou expande. Em ambos casos, existe um *cluster* S' similar a S, porém tem quantidade de elementos a menos ou mais que S. O último bloco de código, após a segunda estrutura de repetição é capaz de identificar quando um novo *cluster* é criado. Tal algoritmo foi executado tanto para o resultado obtido pela clusterização onde a medida de similaridade de Jaccard foi aplicada, quanto para o resultado obtido pela clusterização usando a medida de similaridade *Fading.* Posteriormente os resultados obtidos pela aplicação desta técnica em ambos os casos foram comparados e validados.

Ao fim desta etapa, identificaram-se as transições sofridas pelos assuntos ao longo do tempo, sendo possível observar como tais assuntos repercutiram e se relacionaram.



Algorithm 1: Find each cluster evolution pattern

Input: Similarity matrix M, $C_t, C_{t+\delta t}, \tau, \Delta_{evol}$

```
1  begin
2      if Δ_evol = ∅ then
3          root ← newVertex(root, 0, t)
4          V(Δ_evol) ← V(Δ_evol) ∪ {root}
5      for S_i ∈ C_t do
6          u ← getVertex(Δ_evol, sid_i)
7          if ∃S'_j ∈ C_{t+δt} and M[S_i, S'_j] = 1 then
8              S_i survives
9              v ← newVertex(survives, sid_i, t + δt)
10             V(Δ_evol) ← V(Δ_evol) ∪ {v}
11             E(Δ_evol) ← E(Δ_evol) ∪ {u, v}
12         else
13             if ∃{S'_k ∪ ... ∪ S'_j} ⊆ C_{t+δt}, Σ_{r=k}^{j} M[S_i, S'_r] ≥ τ
               then
14                 S_i splits
15                 for S'_r ∈ {S'_k ∪ ... ∪ S'_j} do
16                     v ← newVertex(split, sid_r, t + δt)
17                     V(Δ_evol) ← V(Δ_evol) ∪ {v}
18                     E(Δ_evol) ← E(Δ_evol) ∪ {u, v}
19             else
20                 if ∃S'_j ∈ C_{t+δt}, {S_i ∪ ... ∪ S_k} ⊆
                   C_t, Σ_{r=i}^{k} M[S_r, S'_j] ≥ τ then
21                     S_i, ..., S_k merge
22                     v ← newVertex(merge, sid_j, t + δt)
23                     V(Δ_evol) ← V(Δ_evol) ∪ {v}
24                     E(Δ_evol) ← E(Δ_evol) ∪ {u, v}
25             else
26                 if ∄S'_j ∈ C_{t+δt}, M[S_i, S'_j] ≥ τ then
27                     S_i disappears
28                     v ← newVertex(disappear, sid_{-1}, t + δt)
29                     V(Δ_evol) ← V(Δ_evol) ∪ {v}
30                     E(Δ_evol) ← E(Δ_evol) ∪ {u, v}
31             else
32                 if ∃S'_j ∈ C_{t+δt}, M[S_i, S'_j] ≥ τ then
33                     if |O_{S_i}| > |O_{S'_j}| then
34                         S_i shrinks
35                         v ← newVertex(shrink, sid_j, t + δt)
36                         V(Δ_evol) ← V(Δ_evol) ∪ {v}
37                         E(Δ_evol) ← E(Δ_evol) ∪ {u, v}
38                     if |O_{S_i}| < |O_{S'_j}| then
39                         S_i expands
40                         v ← newVertex(expand, sid_j, t + δt)
41                         V(Δ_evol) ← V(Δ_evol) ∪ {v}
42                         E(Δ_evol) ← E(Δ_evol) ∪ {u, v}
43     for S'_j ∈ C_{t+δt} do
44         if ∀S_i ∈ C_t, M[S_i, S'_j] = 0 then
45             S'_j appears
46             v ← newVertex(appear, sid_i, t + δt)
47             V(Δ_evol) ← V(Δ_evol) ∪ {v}
48             E(Δ_evol) ← E(Δ_evol) ∪ {root, v}
```

Fonte: Coelho da Silva et al (2014)

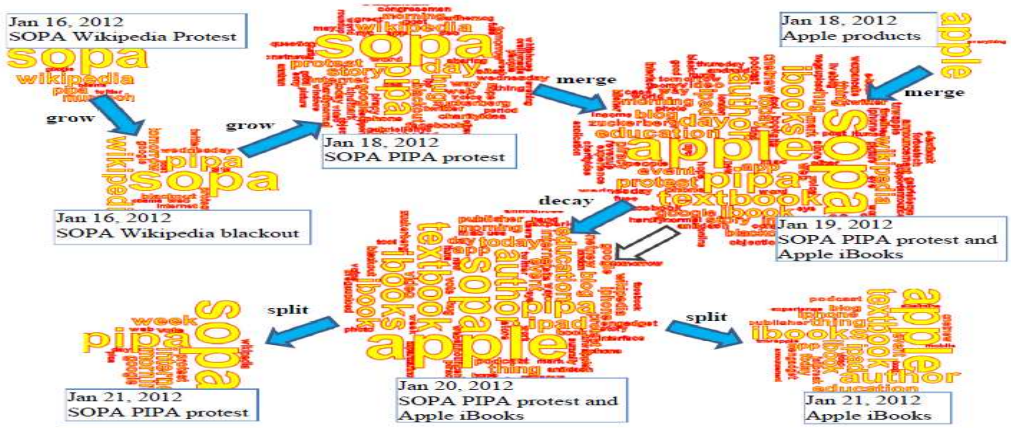
5.5 Elaboração dos gráficos de transição

Após a aplicação das técnicas de evolução de *cluster*, foram gerados gráficos que permitam uma visualização abrangente e panorâmica das transições sofridas pelos *clusters*.

Para representar a dinâmica dos assuntos ao longo do tempo, foram utilizados nos gráficos o recurso de nuvem de palavras. Uma nuvem de palavras é um recurso que permite visualizar os termos mais frequentes de um determinado texto. O tamanho da fonte em que a palavra é apresentada é uma função da frequência da palavra no texto. Desta forma, palavras mais frequentes são escritas em fontes de tamanho maior e palavras menos frequentes são escritas em fontes de tamanho menor.

A Figura 5 ilustra um modelo de gráfico semelhante aos que foram utilizados neste trabalho.

Figura 5- Exemplo de gráfico com nuvens de palavras



Fonte: Lee et al (2014)

5.6 Validação dos resultados obtidos

Esta etapa consiste na constatação dos resultados obtidos por meio de dois passos principais. O primeiro é a comparação dos resultados obtidos pelas duas medidas de similaridade adotadas. Nesse passo, foram analisadas as semelhanças e diferenças apresentadas nas detecções e transições sofridas pelos *clusters* em ambos os casos. No segundo passo, uma análise dos reais acontecimentos ocorridos no período da coleta foi realizada. Através de uma análise subjetiva por meio de noticiários e outras fontes informativas, foi possível observar se os resultados das transições de *clusters* apresentados nos gráficos estão de acordo com o modo que de fato repercutiu determinado assunto naquele período, validando também qual medida de similaridade se aproximou com maior precisão aos fatos reais.

6 RESULTADOS

6.1 Coleta e Pré-Processamento dos dados

A coleta de *tweets* para o trabalho em questão ocorreu de março à junho de 2015. Para a obtenção de um significativo volume de dados a serem analisados, foram selecionados para a coleta quatro assuntos populares e com repercussão nas redes sociais durante o período. Foram eles: terremoto em Nepal, impeachment contra a presidente Dilma, questões referentes ao Movimento Gay e jogos Pan-Americanos.

Para fins de análise, neste trabalho serão estudados apenas os *tweets* referentes aos protestos de impeachment contra a presidenta Dilma e as questões referentes ao Movimento Gay. A coleta feita referente ao terremoto em Nepal e aos jogos Pan-Americanos foi descartada, pois, além da repercussão nas redes sociais ter sido menor se comparado aos outros dois temas, a coleta realizada capturou um grande volume de dados em línguas estrangeiras, o que dificultou a análise dos dados na etapa de clusterização e Evolução de *Clusters* que agrupa os dados com base na semelhança dos textos.

Para coletar dados a respeito dos protestos e questões que giravam em torno da presidenta do Brasil, foram capturados *tweets* que contivessem a palavra “Dilma” durante todo o mês de março. Considerando o vultuoso volume de dados entre 08 de março e 16 de março de 2015, os dados analisados foram referente a este período. Da mesma maneira, para a captura de *tweets* sobre as questões referentes ao Movimento Gay repercutidas no período da coleta, foi utilizado como parâmetro a palavra “Gay”, tendo em vista que essa era palavra que marcava temas como “Parada Gay” e “Casamento Gay”, assuntos muito debatidos nas redes sociais no primeiro semestre de 2015. A coleta com o termo “Gay” se deu durante todo o mês de junho, sendo considerado para análise o período compreendido entre 03 de junho à 12 de junho de 2015, período este em que o volume de *tweets* sobre o assunto foi consideravelmente maior. Tais palavras foram escolhidas por serem imparciais e abrangentes, possibilitando uma ampla coleta dos temas em questão.

Após a fase de coleta, por meio da técnica de remoção de *stopwords* anteriormente explicada na seção 3.2.1, foram removidos de todos os *tweets* coletados conteúdos não relevantes semanticamente para a fase de clusterização: links, nomes de usuário do Twitter, artigos, preposições, caracteres especiais, caracteres numéricos, dentre outros. As palavras “Dilma” e “Gay”, por terem sido utilizadas para a captura dos dados e estarem presente em todos os *tweets* de suas respectivas coletas, também foram removidas dos *tweets*. Desta forma,

após o processamento restaram apenas as palavras chaves na composição do *tweet*, contribuindo de forma positiva para as próximas etapas de análise dos dados.

A Tabela 2 a seguir apresenta informações referentes aos dados coletados ao longo dos dias para os respectivos temas analisados.

Tabela 3 - Resultado da Coleta de Dados

Tema	Período da Coleta Analisado	Qtd de *tweets*	Tamanho dos dados coletados	Tamanho dos dados processados
Dilma	08/03/15 à 16/03/15	672.551	512 Mb	327 Mb
Gay	03/06/15 à 12/06/15	798.362	619 Mb	386 Mb

6.2 Clusterização dos dados

Para a clusterização dos dados, os *tweets* coletados e processados foram decompostos em uma série de *snapshots* ao longo dos dias da coleta. Os seguintes parâmetros foram repassados para o algoritmo de clusterização:

- Tipo de entrada de dados: CSV
- *eps*: 0.3
- *minPts*: 0.25 * (quantidade de *tweets* do *snapshot*)

No contexto deste trabalho o *eps* é um valor limite que mede a similaridade de dois *tweets*. Logo, o resultado obtido após a aplicação da medida de similaridade entre dois *tweets*, deve ser igual ou maior que 0.3 para que sejam considerados similares. Já o *minPts* é o parâmetro que define a densidade de um *cluster.* Dessa forma, um *cluster* agrupa pelo menos 25% dos *tweets* do *snapshot*. Tais valores foram escolhidos com base nos valores utilizados em LEE et al (2014) e validados por meio de experimentação nos *snapshot*s do trabalho em questão, utilizando outros valores como teste para *eps* e *minPts* e constatando que os valores apresentados em LEE et al (2014) favoreciam uma detecção mais abrangente de *clusters.*

O arquivo CSV passado para o algoritmo como entrada, corresponde a um *snapshot* e contém um dia de coleta de dados. Essa etapa de clusterização ocorreu duas vezes para cada *snapshot*, uma empregando a medida de similaridade de Jaccard e outra empregando a medida de similaridade *Fading*. Para fins de aplicação da medida *Fading*, a distância de tempo entre um *tweet* e outro foi contabilizada em horas.

Como resultado da clusterização de um *snapshot*, a saída obtida foi:

- k *clusters* (um arquivo CSV para cada *cluster* detectado, contendo seus respectivos *tweets*, cada *tweet* marcado como *core* ou *border* (definidos pelo algoritmo DBSCAN));

Realizada a clusterização dos dados e obtidos os *clusters* dos assuntos repercutidos, foi possível passar para a próxima etapa e aplicar o algoritmo de Evolução de *Clusters*.

6.3 Evolução de *Clusters*

Nesta seção serão apresentados todos os resultados obtidos por meio do emprego da técnica de Evolução de *Clusters* sobre os dados clusterizados. Os resultados serão apresentados referentes aos dois temas analisados, no qual o Tema 1 se trata da coleta realizada sobre a presidenta Dilma Roussef e o Tema 2 se refere aos eventos e assuntos detectados a respeito do movimento Gay.

Com base nos valores adotados em LEE et al (2014),o parâmetro τ (*threshold* de similaridade) passado para o algoritmo de evolução foi de 0.3. Esse parâmetro é utilizado como base para comparar os pontos cores dos *clusters* em momentos distintos e detectar possíveis evoluções. Como explicado melhor na seção 4.4, foram passados também como entrada do algoritmo os parâmetros C_t e $C_{t+\delta t}$, que são todos os *clusters* encontrados no *snapshot* t e todos os *clusters* encontrados no *snapshot* t + δt. Além disso, também foi repassada a matriz M para armazenar em cada célula o valor de similaridade entre os *clusters* de *snapshot*s diferentes.

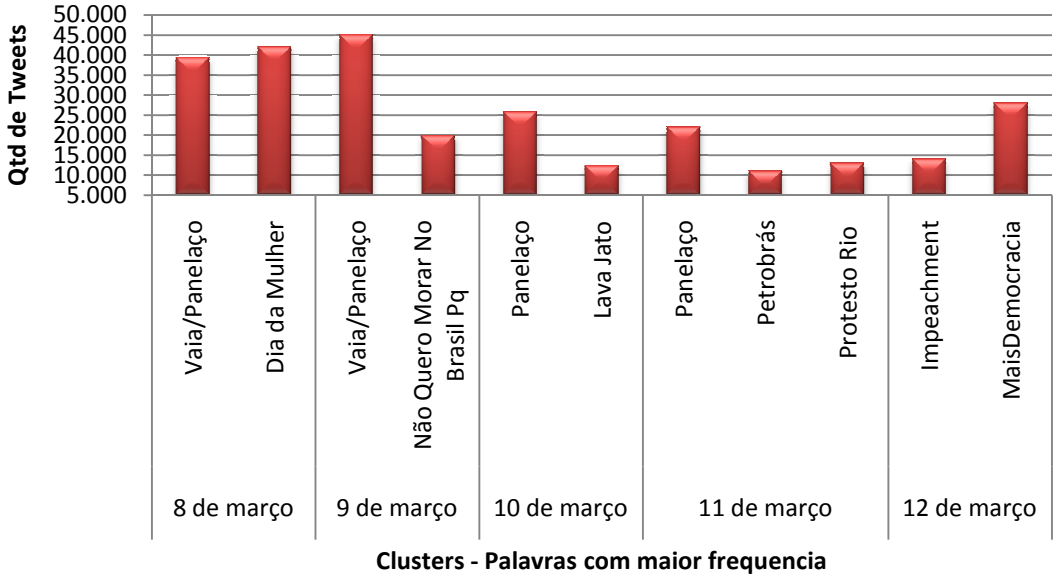
Na seção 5.3.1 são mostrados os resultados da clusterização utilizando a medida de similaridade de Jaccard e a posterior evolução dos *clusters* identificados para o Tema 1. A seção 5.3.2 apresenta os resultados obtidos na clusterização utilizando a medida de similaridade *Fading* e a evolução dos *clusters* identificados utilizando esta medida. A seguir, na seção 5.3.3 é apresentada a análise do Tema 2 utilizando a medida de similaridade de Jaccard. Por fim, a seção 5.3.4 expõe os resultados das evoluções de *clusters* do Tema 2 com a medida de similaridade *Fading*. Em todas essas seções, são referenciados noticiários que confirmam os resultados apresentados.

6.3.1 Análise do Tema 1 com a medida de similaridade de Jaccard

Os resultados apresentados nesta seção referem-se ao emprego do algoritmo de Evolução de *Clusters*, sobre os *clusters* gerados a partir do emprego do *DBSCAN* com a medida de

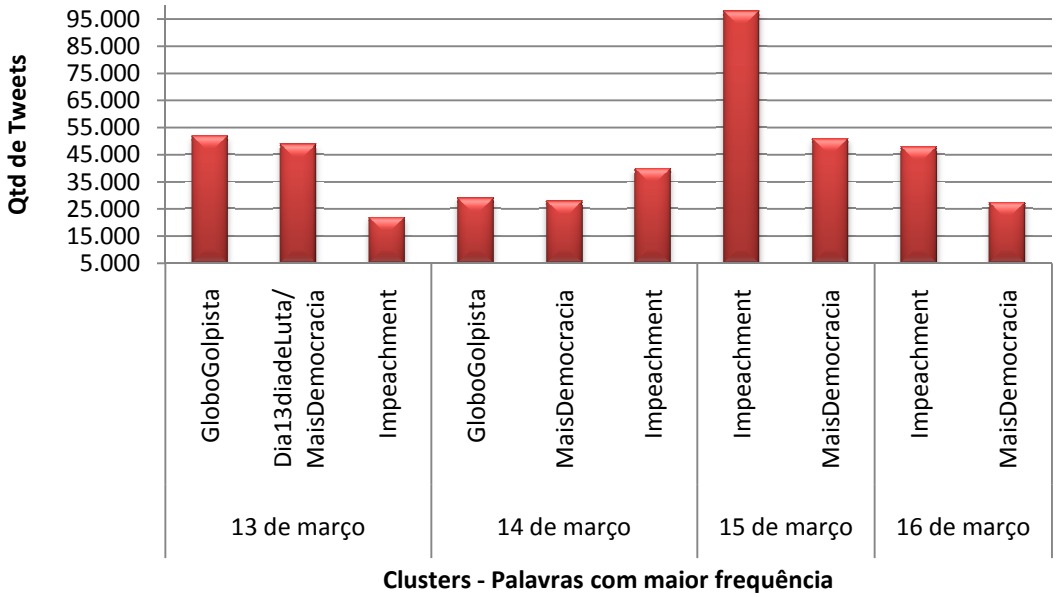
similaridade de Jaccard, nos dados coletados com o termo “Dilma”. As Figuras 6 e 7 apresentam, em linha temporal, os *clusters* detectados ao longo dos dias de análise.

Figura 6- *Clusters* gerados com a medida de Jaccard sobre o Tema 1 (08/03/15 à 12/03/15)



Fonte: Elaborado pelo Autor

Figura 7- *Clusters* gerados com a medida de Jaccard sobre o Tema 1 (13/03/15 à 16/03/15)



Fonte: Elaborado pelo Autor

No dia 08 de março foram detectados dois assuntos: um sobre o Dia da Mulher, data celebrada neste dia, onde a maioria dos *tweets* deste *cluster* continham a *hashtag*

#DilmadaMulher. E outro assunto, marcado pela *hashtag* #VaiaDilma, que se referia a vaia que ocorreu no momento do pronunciamento realizado pela presidenta para as mulheres naquele mesmo dia[5]. Tais assuntos polarizaram apoiadores e críticos do governo petista. No dia seguinte, os dois assuntos passaram a repercutir juntos em um mesmo *cluster* que se referia ao "panelaço" (como ficou conhecido o protesto realizado no momento do pronunciamento no Dia da Mulher)[6], sendo detectado pelo algoritmo **uma união de *clusters* como evolução**. O assunto marcado pela *hashtag* "#NãoQueroMorarNoBrasilpq" também foi **um *cluster* que apareceu** dia 09 de março no Twitter, mas no dia seguinte **desapareceu**. No dia 10 de março, os *tweets* a respeito do panelaço passaram a diminuir, fazendo com que o ***cluster* sofresse uma retração**. Ainda no dia 10 de março, **surgiu um *cluster*** que levantava debates sobre a operação Lava Jato mas que também **desapareceu** no dia seguinte. Em 11 de março, **o *cluster* sobre o panelaço continuou a diminuir**. Outros dois ***clusters*** foram **detectados (aparecimento)**, um sobre a Petrobrás e outro que se tratava sobre um pequeno protesto ocorrido no Rio contra a presidente[7]. No dia 12 de março, o *cluster* que outrora repercutia sobre o panelaço **dividiu-se em dois outros *clusters***, um de apoiadores do governo Dilma que levantava a *hashtag* "menosOdioMaisDemocracia" e outro com teor opositor ao governo e que repercutia pedidos de impeachment. O *cluster* que repercutia favorável ao governo possuía maior densidade de *tweets* por fazer menção a uma manifestação que estava sendo organizada para o dia seguinte por apoiadores do governo Dilma. No dia 13 de março, a discussão que repercutia a favor do governo **sofreu ainda uma maior expansão** repercutindo a *hashtag* "#Dia13Diadeluta", tendo em vista que nesse mesmo dia estava ocorrendo em todo o país uma manifestação organizada pela CUT (Central Única dos Trabalhadores)[8]. Tal manifestação ocorreu em defesa da Petrobrás, dos direitos e da Reforma Política apoiada pelo governo Dilma. No mesmo dia, foi detectado **o aparecimento de um *cluster*** que repercutia em torno da *hashtag* "#globoGolpista", no qual simpatizantes da manifestação a favor do governo alegavam que a emissora de televisão rede Globo estaria desfavorecendo as manifestações do dia 13 de março ao omitir sua proporção. Ainda no dia 13, **o *cluster*** que repercutia com pedidos de impeachment sofreu uma pequena **expansão**. Já no dia 14 de março, esse mesmo *cluster* favorável ao impeachment **continuou a expandir**, ao passo que o

[5] http://g1.globo.com/politica/noticia/2015/03/pessoas-protestam-durante-pronunciamento-de-dilma.html
[6] http://www1.folha.uol.com.br/poder/2015/03/1600073-em-cidades-com-panelaco-internautas-tambem-defendem-dilma.shtml
[7] http://www1.folha.uol.com.br/poder/2015/03/1601360-ato-contra-dilma-reune-15-manifestantes-e-150-pms-na-candelaria.shtml
[8] http://www.cut.org.br/noticias/cut-mobilizada-para-o-dia-13-de-marco-664c/

cluster positivo ao governo que repercutia a *hashtag* “MenosOdioMaisDemocracia” e o *cluster* que repercutia negativamente contra a emissora rede Globo **sofreram retração**. Dia 14 era véspera do protesto organizado nacionalmente contra a presidenta Dilma Roussef. No dia 15 de março, dia do protesto, a repercussão em torno das manifestações a favor do impeachment sofreu **grande expansão**. O *cluster* favorável ao governo, embora tenha expandido em comparação com o dia anterior, não alcançou a expansão do *cluster* a favor do impeachment. O ***cluster*** negativo a Rede Globo **desapareceu** no dia 15. Nos dias seguintes a este evento, os **dois assuntos sofreram retração** e continuaram a evoluir de forma independente.

A Figura 8 ilustra, por meio do algoritmo de evolução empregado, as transições sofridas pelos *clusters* detectados com teor positivo e negativo ao governo Dilma. Note que foi possível capturar a evolução dos assuntos em formato de *clusters*, de acordo com os objetivos deste trabalho.

Figura 8- Evolução dos assuntos sobre "Dilma" com Jaccard entre 08/03/15 e 15/03/15

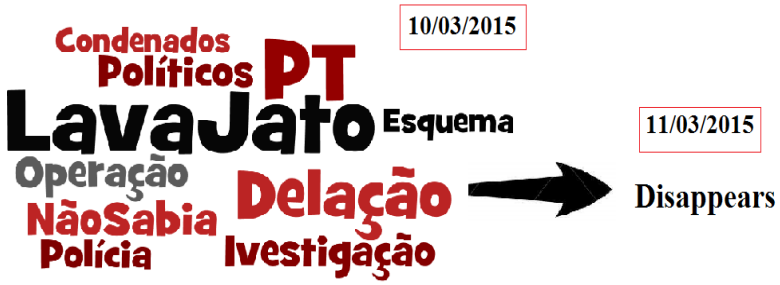
Fonte: Elaborado pelo Autor.

As Figuras 9, 10, 11, 12 e 13 a seguir, apresentam as nuvens de palavras dos *clusters* que surgiram e desaparecerem ao longo dos dias, conforme cita.

Figura 9: Evolução de *Cluster* detectado em 09/03/2015

Fonte: Elaborado pelo Autor.

Figura 10: Evolução de *Cluster* detectado em 10/03/2015

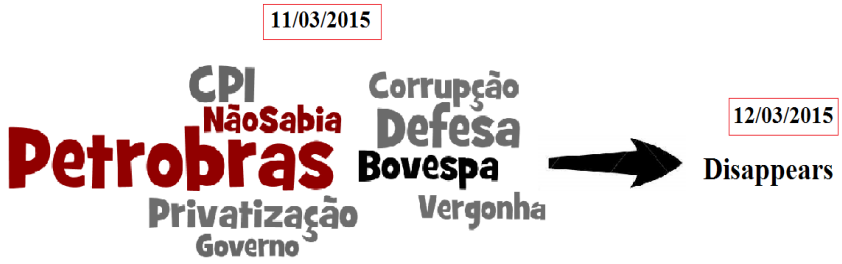


Fonte: Elaborado pelo Autor.

Figura 11: Evolução de *Cluster* detectado em 11/03/2015

Fonte: Elaborado pelo Autor.

Figura 12: Evolução de *Cluster* detectado em 11/03/2015



Fonte: Elaborado pelo Autor.

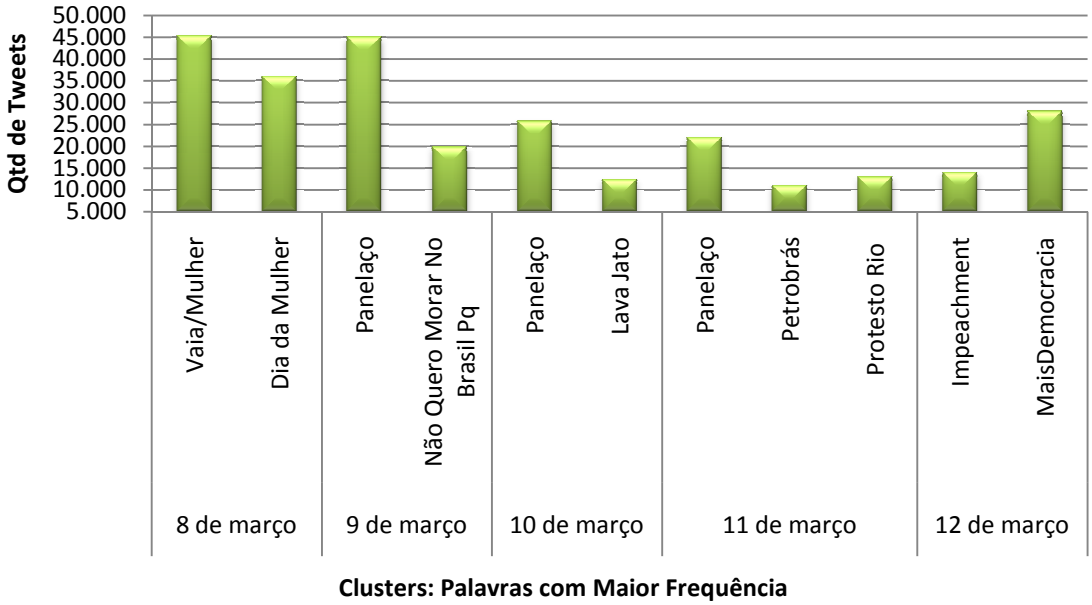
Figura 13: Evolução de *Cluster* detectado em 13/03/2015

Fonte: Elaborado pelo Autor.

6.3.2 Análise do Tema 1 com a medida de similaridade *Fading*

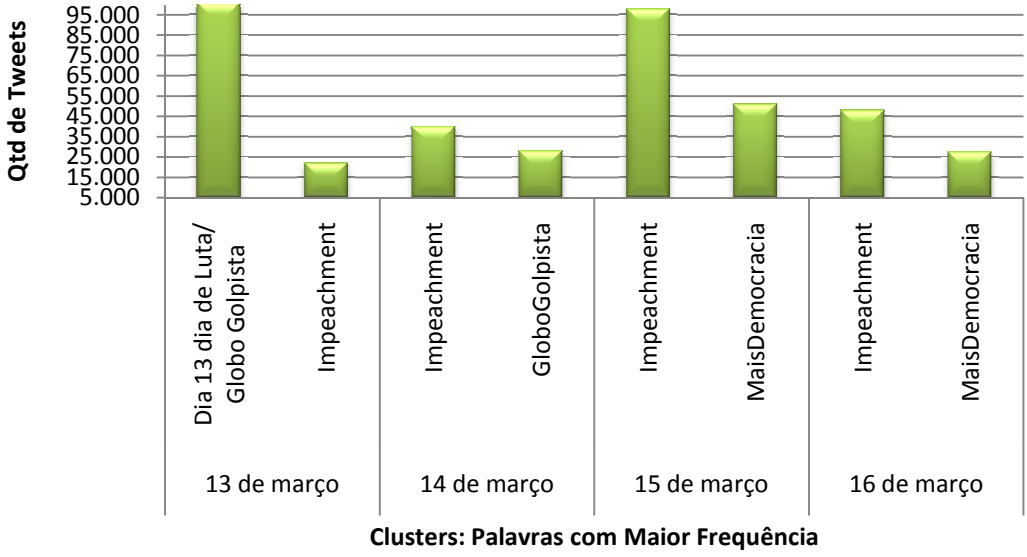
Para a mesma base de dados utilizada na análise anterior, subseção 5.3.1, foi empregada a clusterização com a medida de similaridade *Fading* e posteriormente o algoritmo de evolução para detectar as transições sofridas pelos assuntos identificados na clusterização. As Figuras 14 e 15 apresentam os *clusters* identificados utilizando a medida *Fading* ao longo dos dias 08 à 16 de março de 2015.

Figura 14- *Clusters* gerados com a medida de Fading sobre o Tema 1 (08/03/15 à 12/03/15)



Fonte: Elaborado pelo Autor

Figura 15- *Clusters* gerados com a medida de Fading sobre o Tema 1 (13/03/15 à 16/03/15)



Fonte: Elaborado pelo Autor

Após a análise foi possível observar que o resultado da clusterização com a medida de similaridade *Fading* diferiu do resultado utilizando a medida de Jaccard, mais especificamente, em dois dias: 08 e 13 de março.

Assim como na análise com Jaccard, no dia 08 de março foram detectados 2 *clusters*: um que se referia ao Dia da Mulher e outro que se referia a "vaia" organizada para a presidenta

Dilma Roussef durante o seu pronunciamento em rede nacional felicitando as mulheres pelo seu dia. No entanto, observou-se que a densidade do *cluster* sobre o Dia da Mulher foi consideravelmente menor que no resultado da clusterização com Jaccard (o *cluster* aplicando a medida de similaridade de Jaccard continha 39.345 *tweets*, com a medida *Fading* passou a possuir 45.492 *tweets*). Para entender as motivações de tais mudanças, foi analisado o horário em que os *tweets* foram publicados. Com isso foi possível constatar que mais de 70% dos *tweets* pertencentes ao *cluster* sobre a Vaia foram publicados à noite, período em que a presidenta Dilma fez o pronunciamento. Utilizando a clusterização com *Fading,* os *tweets* que ficaram no *cluster* sobre o "Dia da Mulher", embora marcados pela *hashtag* "DilmaDaMulher", eram *tweets* que prestavam homenagem as mulheres de uma forma geral e as parabenizavam pelo seu dia. Enquanto praticamente todos os *tweets* que se tratavam em específico sobre a "vaia" ocorrida por ocasião do pronunciamento da presidenta no Dia da Mulher, foram corretamente alocados para o *cluster* que se referia sobre a "vaia", o que não ocorreu na clusterização com Jaccard. Usando Jaccard, muitos *tweets* publicados durante a noite e que se referiam as vaias ocorridas e ao pronunciamento no "Dia da Mulher", foram alocados no *cluster* sobre o "Dia da Mulher", no entanto, com *Fading,* foram alocados no *cluster* mais específico sobre este assunto que era o que tratava sobre o Panelaço e a Vaia ocorrida. Desta forma, ao considerar a distância no tempo em que os *tweets* foram publicados, o algoritmo utilizando a medida *Fading* pôde eficazmente aproximar *tweets* cronologicamente pertos e fazer uma distribuição de *tweets* de forma mais precisa. Não apenas aproximando os *tweets* por assunto, mas pelo instante em que foram publicados. A seguir, nas Figuras 16 e 17 que apresentam as nuvens de palavras do *cluster* "Dia da Mulher" com Jaccard e com a medida *Fading*, é possível observar que palavras como "pronunciamento", "democracia", "BrasilTeAma" e "força" deixam de fazer parte de um *cluster* para outro. Isso por que, a maioria dessas palavras foi publicada em *tweets* que defendiam a presidenta no momento do pronunciamento.

Figura 176 - Cluster "Dia da Mulher" com Jaccard

Fonte: Elaborado pelo Autor.

Figura 16 - Cluster "Dia da Mulher" com *Fading*

Fonte: Elaborado pelo Autor.

Além disso, no dia 09 de março, o *cluster* referente a "vaia" **sofreu uma expansão** e o *cluster* sobre o dia da Mulher **desapareceu**, diferente do resultado anterior onde os dois *clusters* sofriam uma união.

Nos dias 10 e 11 de março o ***cluster*** sobre o Panelaço **passou a sofrer retrações**, até passar por uma **divisão** no dia 12 que distinguiu a repercussão de apoiadores e opositores do governo, da mesma forma como ocorreu com a medida de Jaccard.

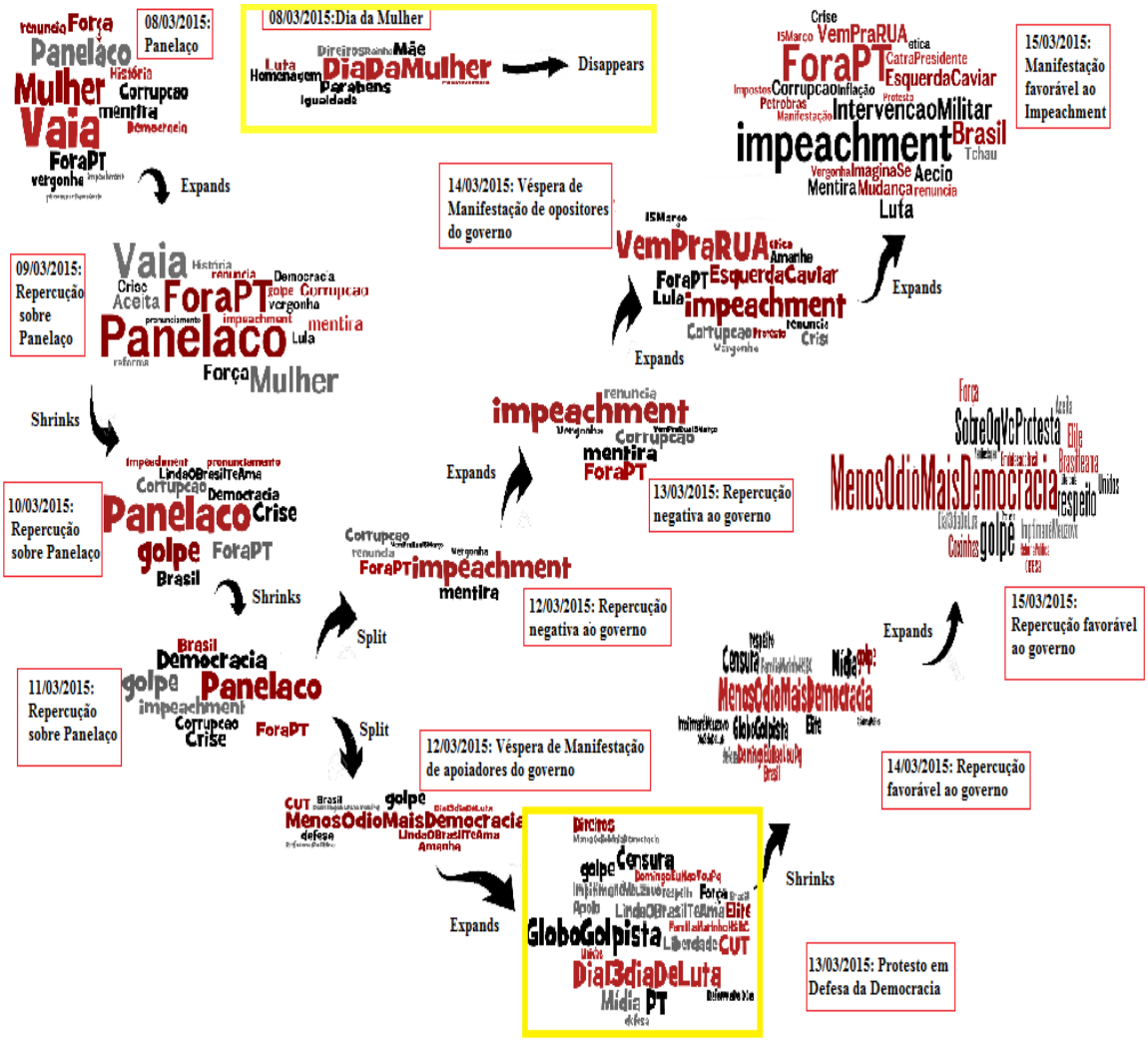
No dia 13 de março, dia em que ocorria um protesto em defesa do governo, enquanto o algoritmo com a medida de Jaccard detectou os assuntos "Dia 13 dia de Luta" e "Globo Golpista" em *clusters* distintos, o algoritmo com a medida *Fading* detectou ambos os temas em um mesmo *cluster*, decorrentes da **expansão** do *cluster* que apoiava o governo**.** De fato, tais assuntos estavam estreitamente relacionados, tendo em vista que os *tweets* direcionados de forma negativa a rede Globo defendiam que a emissora estaria desfavorecendo o governo da presidente Dilma Rousseff com manchetes tendenciosas, além de citar a omissão da emissora quanto aos protestos que estavam acontecendo no dia 13 de março em prol do governo. Os manifestantes chegaram a bradar a frase "Abaixo a rede Globo", na Avenida Paulista, em São Paulo[9]. Desta forma, é possível afirmar que os assuntos "Dia 13 dia de Luta" e "Globo Golpista" realmente faziam parte de um mesmo *cluster*, e pôde ser eficazmente detectado pelo algoritmo com a medida *Fading*, que ao considerar a proximidade no tempo com que tais *tweets* estavam sendo publicados os alocou em um só *cluster* identificando sua

[9] http://popxd.com/2015/03/globo-golpista.html

similaridade. Os demais *clusters* dos dias 14 à 16 de março seguiram o mesmo padrão de evolução detectado pela medida de similaridade de Jaccard.

A seguir, na Figura 18, é possível observar a evolução dos *clusters* detectados com a medida de similaridade *Fading*.

Figura 18- Evolução dos assuntos sobre "Dilma" com Fading entre 08/03/15 e 15/03/15

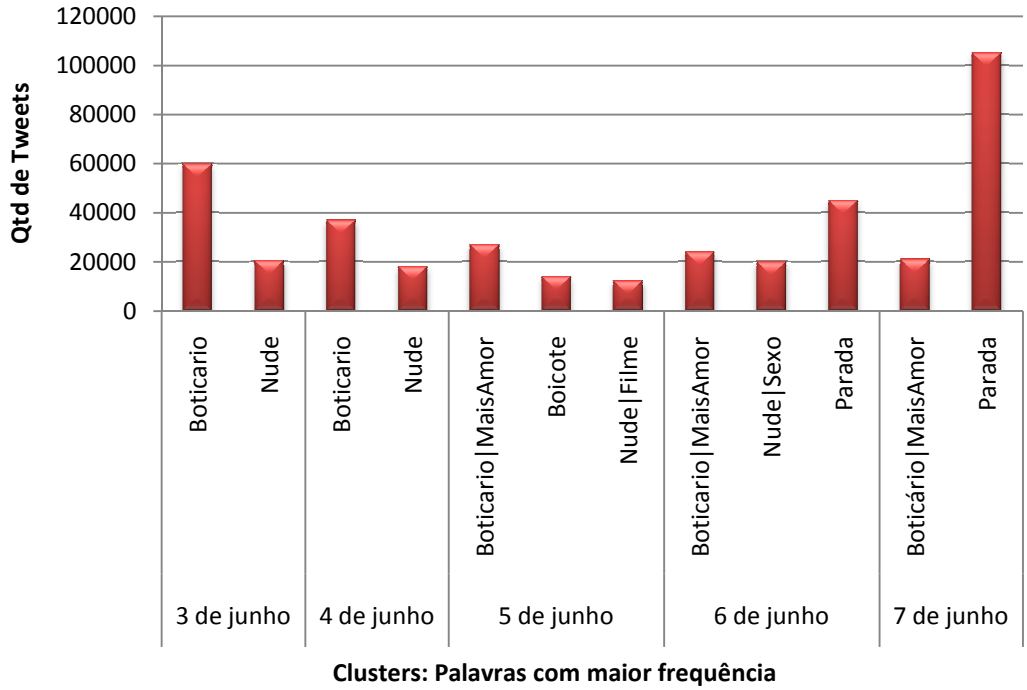


Fonte: Elaborado pelo Autor.

6.3.3 Análise do Tema 2 com a medida de similaridade de Jaccard

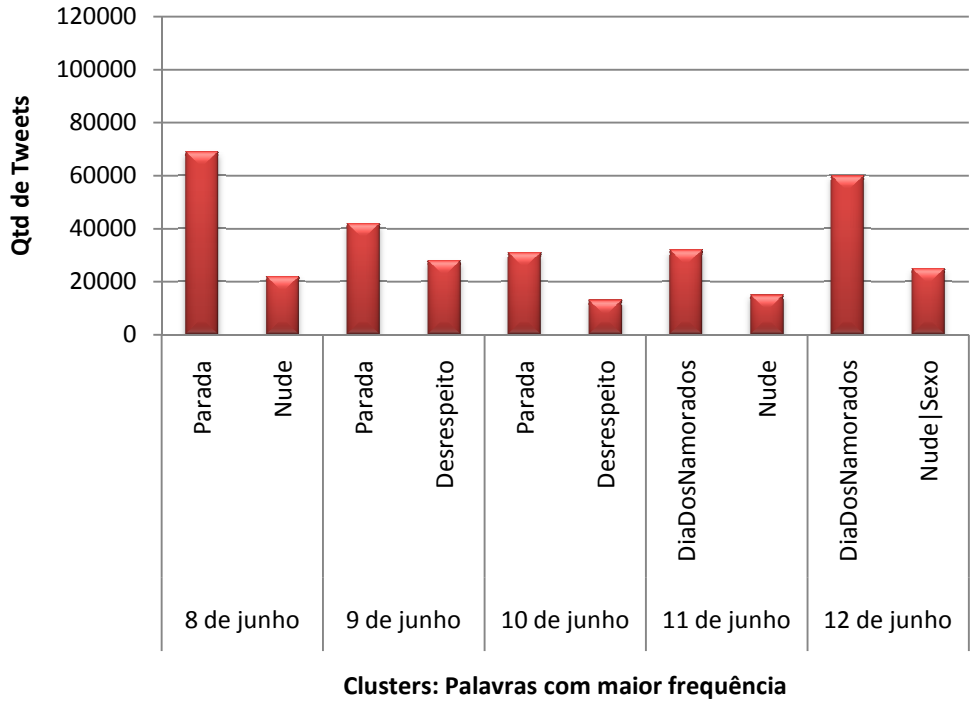
Os resultados apresentados nesta seção referem-se ao emprego do algoritmo de Evolução de *Clusters*, sobre os *clusters* gerados a partir do emprego do *DBSCAN* com a medida de similaridade de Jaccard, nos dados coletados com o termo “Gay” entre 03/06/2015 e 12/06/2015. As Figuras 19 e 20 apresentam, em linha temporal, os *clusters* detectados ao longo dos dias de análise.

Figura 19 - *Clusters* gerados com a medida de Jaccard sobre o Tema 2 (03/06/15 à 07/06/15)



Fonte: Elaborado pelo Autor.

Figura 20 - *Clusters* gerados com a medida de Jaccard sobre o Tema 2 (08/06/15 à 12/06/15)



Fonte: Elaborado pelo Autor.

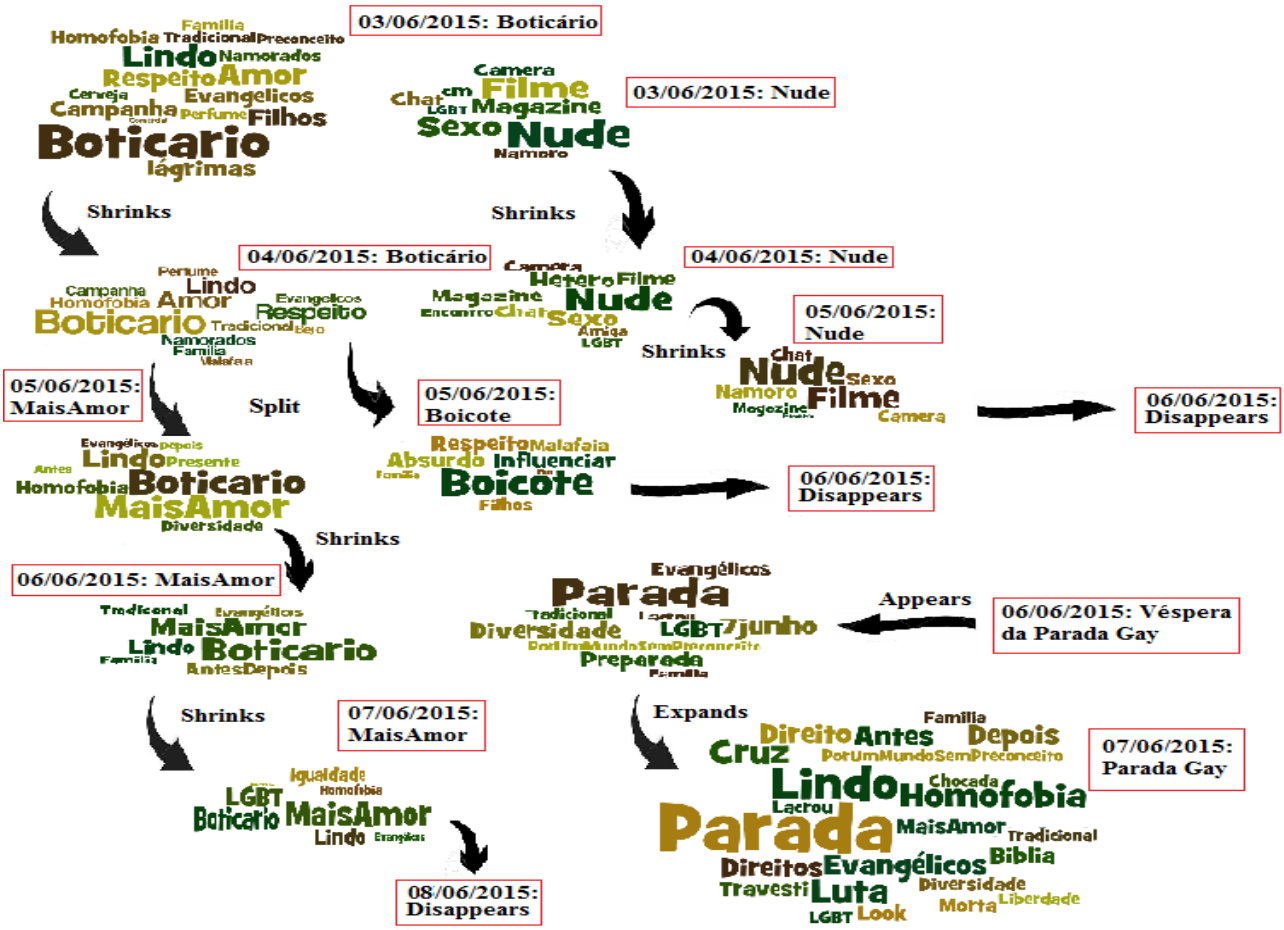
No primeiro dia de análise, 03 de junho de 2015, dois *clusters* foram detectados. O mais expressivo deles em quantitativo de *tweets*, se tratava da Campanha de dia dos Namorados da

marca de cosméticos OBoticário. A campanha destacava uma fragrância unissex, o Egeo, apresentando casais gays trocando presentes[10]. A propaganda gerou grande repercussão e polêmica desde sua divulgação em 25 de maio de 2015. O outro assunto detectado no mesmo dia apresentava teor pornográfico, com destaque para a palavra "Nude" no *cluster*. No dia seguinte, **ambos os *clusters* sofreram retração**. Em 05 de junho de 2015, o *cluster* sobre a Campanha de OBoticário **sofreu uma divisão**, onde o maior *cluster* resultante dessa divisão era marcado por palavras positivas em relação a campanha, enquanto o *cluster* de menor expressividade denotava teor crítico e opositor, onde os usuários incentivavam a não aquisição de produtos OBoticário, gerando ameaças de boicote à marca[11]. O *cluster* com teor pornográfico se manteve pelo terceiro dia consecutivo de análise, sofrendo **retrações**. Em 06 de junho, o ***cluster* negativo** a campanha OBoticário havia **desaparecido** e o ***cluster* positivo** a campanha manteve-se sofrendo uma **retração**. Neste mesmo dia, foi detectado também o **aparecimento de um *cluster*** referente a 19ª Parada do Orgulho LGBT (Lésbicas, Gays, Bissexuais, Travestis e Transexuais) que ocorreria em São Paulo no dia seguinte. No dia 07 de junho, **o *cluster*** referente a Parada Gay sofreu uma **expansão**, tendo em vista que este foi o dia em que o evento ocorreu[12]. Palavras como "Homofobia" e "Lindo" foram destaques neste *cluster*. A repercussão positiva frente a campanha da marca OBoticário manteve-se em 07 de junho, sofrendo **retração**. Em 08 de junho, **o *cluster*** que se mantinha repercutindo sobre a campanha de dia dos namorados OBoticário **desapareceu**. No mesmo dia, **o *cluster*** sobre a ParadaGay sofreu **retração** e foi **detectado novamente o *cluster*** que continha teor pornográfico. No dia 09 de junho, o assunto referente a Parada Gay passou a **dividir-se** entre uma vertente de defensores do movimento, com *tweets* positivos e de incentivo ao evento, e uma vertente que fez duras críticas e desaprovou a Parada Gay considerando-a desrespeitosa, principalmente pelo fato de uma Transexual ter se vestido como Jesus Cristo e encenado a própria crucificação durante o movimento[13]. Segundo Viviany Beleboni, a transexual que protagonizou tal cena, a representação foi apenas uma metáfora e fez referência a todas as mortes e agressões contra a comunidade gay. No dia seguinte **ambos os *clusters* sofreram retração**, no entanto, o *cluster* positivo permaneceu com maior expressividade em número de

[10] http://economia.uol.com.br/noticias/redacao/2015/05/26/o-boticario-usa-gays-em-campanha-do-dia-dos-namorados-reveja-outras.htm
[11] http://g1.globo.com/economia/midia-e-marketing/noticia/2015/06/comercial-de-o-boticario-com-casais-gays-gera-polemica-e-chega-ao-conar.html
[12] http://g1.globo.com/sao-paulo/noticia/2015/06/parada-gay-reune-milhares-em-sp.html
[13] http://www.jornalopcao.com.br/ultimas-noticias/transexual-crucificada-na-parada-gay-de-sp-vira-alvo-de-polemica-nas-redes-sociais-37593/

tweets que o *cluster* negativo ao movimento. No dia 11 de junho, véspera do dia dos namorados, o ***cluster* positivo** a Parada Gay passou a ter muitos *tweets* se referindo a esta data, sofrendo assim uma **expansão**, ao passo que o ***cluster* negativo** sobre a Parada Gay **desapareceu**. O ***cluster*** com conteúdo pornográfico **voltou a ser detectado** no dia 11 de junho. No dia 12 de junho, dia dos namorados, **ambos os *clusters*** detectados no dia 11 sofreram **expansão**, embora o *cluster* referente ao dia dos namorados tenha abrangido um quantitativo de *tweets* expressivamente maior. É importante ressaltar que o *cluster* referente ao dia dos namorados teve dentre as palavras destaque o termo "Casamento", no qual muitos *tweets* traziam reivindicações e apoio em prol da aprovação do casamento Gay, posteriormente legalizado pelos Estados Unidos no dia 26/06/2015[14]. A análise foi concluída no dia 12 de junho, com 10 dias de coleta de dados sobre esse tema. Nas Figuras 21 e 22 é possível observar, por meio das nuvens de palavras, as evoluções dos assuntos detectados.

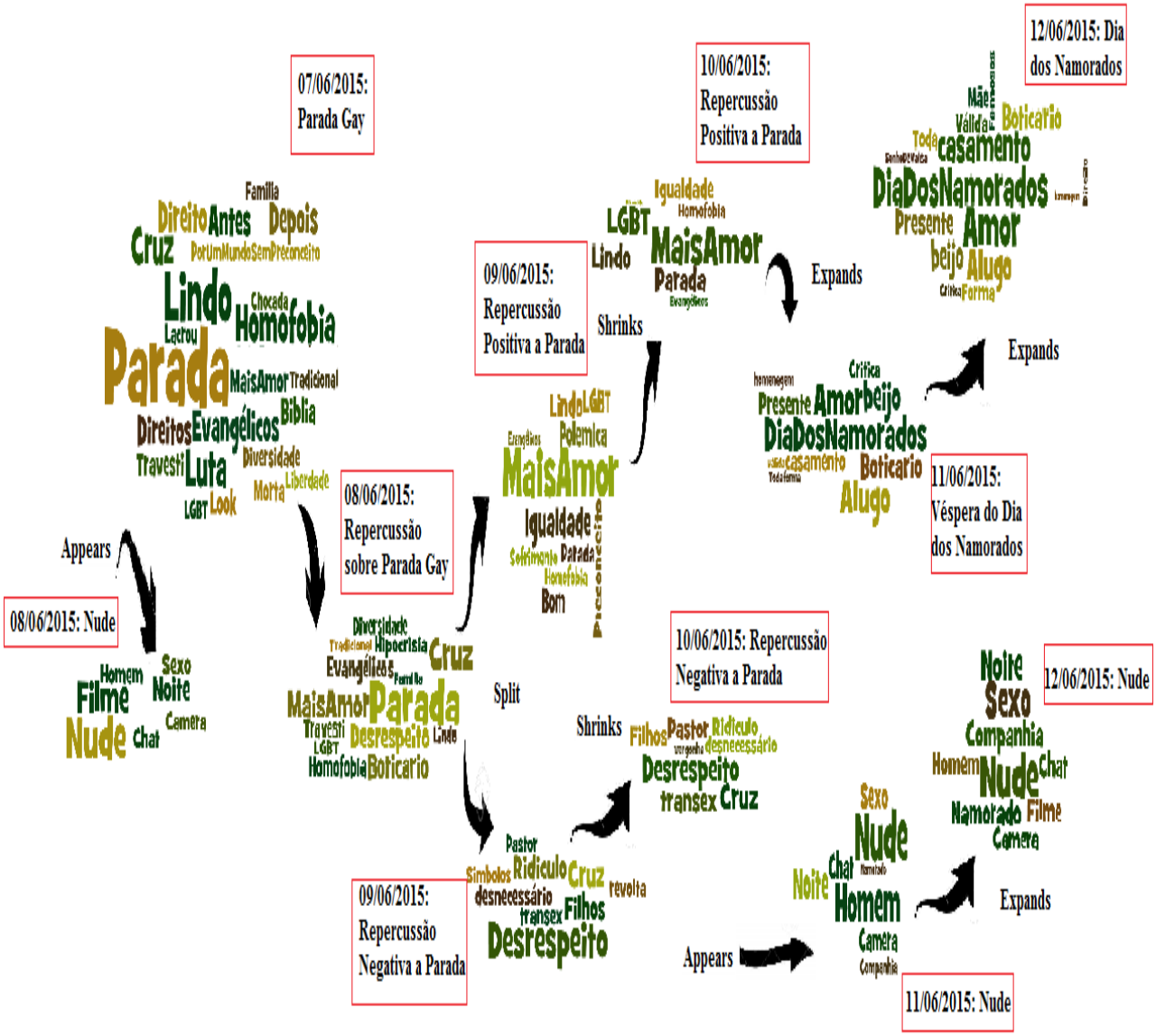
Figura 21- Evolução dos assuntos sobre o tema "Gay" com Jaccard entre 03/06/15 e 07/06/15



Fonte: Elaborado pelo Autor.

[14] http://www.bbc.com/portuguese/noticias/2015/06/150622_casamento_gay_legaliza_eua_rb

Figura 22 - Evolução dos assuntos sobre o tema "Gay" com Jaccard entre 07/06/15 e 12/06/15

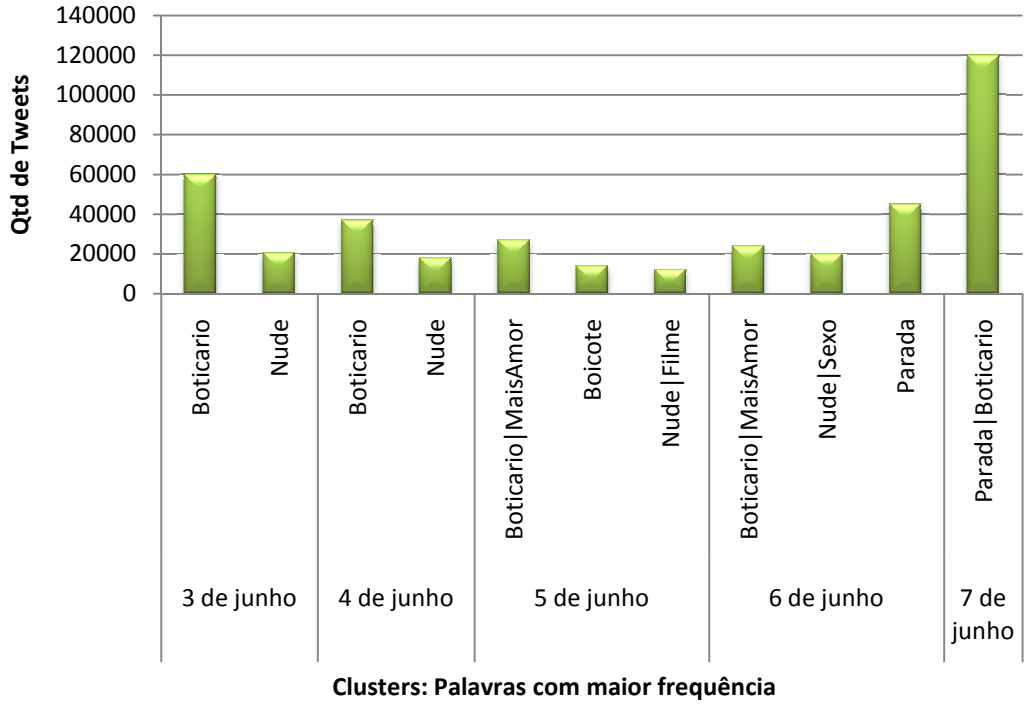


Fonte: Elaborado pelo Autor.

6.3.4 Análise do Tema 2 com a medida de similaridade de *Fading*

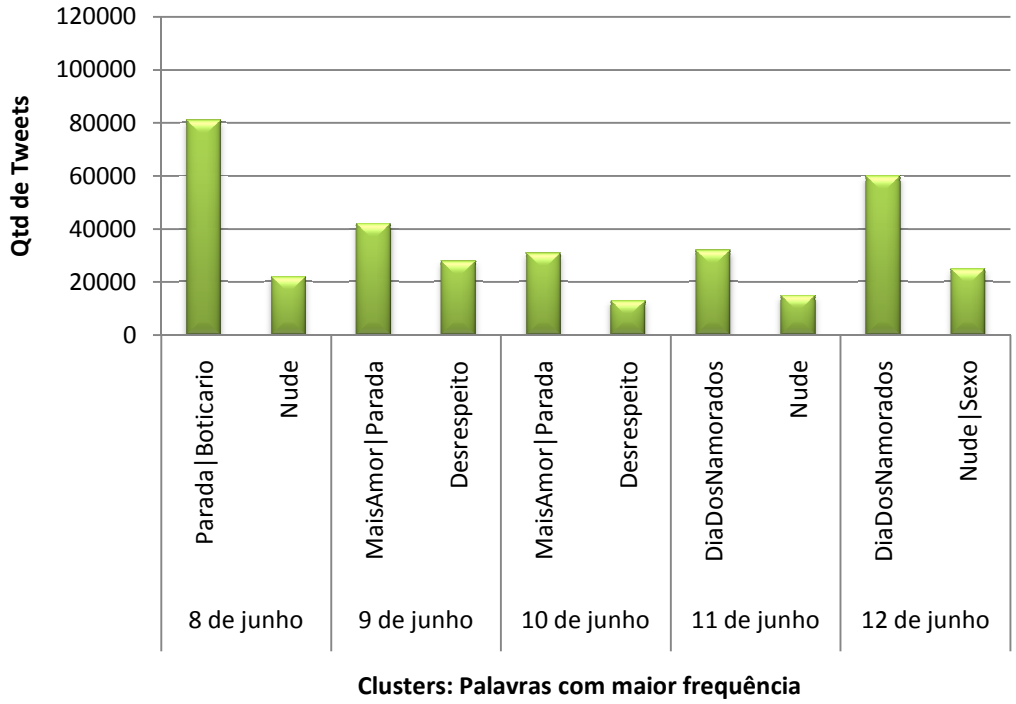
Utilizando a mesma base de dados empregada na análise com Jaccard na subseção anterior, foi empregada a clusterização com a medida de similaridade *Fading* e posteriormente o algoritmo de evolução para detectar as transições sofridas pelos assuntos identificados na clusterização. É possível observar nas Figuras 23 e 24 a seguir, os *clusters* identificados utilizando a medida *Fading* ao longo dos dias 03 a 12 de junho de 2015.

Figura 23 - *Clusters* gerados com a medida *Fading* sobre o Tema 2 (03/06/15 à 07/06/15)



Fonte: Elaborado pelo Autor.

Figura 24 - *Clusters* gerados com a medida *Fading* sobre o Tema 2 (07/06/15 à 12/06/15)



Fonte: Elaborado pelo Autor.

Foi constatado que o resultado da aplicação da medida de similaridade *Fading* no Tema 2, difere do resultado com a medida de Jaccard nos dias 07 e 08 de Junho.

Dos dias 03 a 06 de junho, os *clusters* detectados seguiram o padrão da análise com Jaccard, sendo possível observar a repercussão sobre a campanha de dia dos namorados OBoticário e a repercussão sobre o conteúdo relacionado a pornografia Gay. Ambos os temas, detectados no dia 03 de junho, **sofreram retração** no dia 04. No dia 05, o assunto sobre a campanha de OBoticário **se dividiu** em uma corrente que repercutia positivamente e outra que atacou a campanha com críticas e ameaças de boicote a marca. Assim como com Jaccard, no dia 06 foram detectados 3 assuntos: um sobre a repercussão positiva referente a campanha de OBoticário (ao passo que a repercussão negativa desapareceu), outro sobre pornografia Gay e por fim, um terceiro que se tratava da Parada Gay que ocorreria no dia seguinte.

No dia 07 de junho foi possível observar diferenças entre a transição apresentada na análise com a medida *Fading* em detrimento da análise onde se aplicou Jaccard. Com a medida *Fading* foi detectado no dia 07 apenas um expressivo *cluster*, que se referia de maneira geral a Parada Gay (evento que ocorreu neste dia em São Paulo). Tal *cluster* foi resultado da **união** entre a repercussão positiva da campanha de OBoticário e o *cluster* sobre a Parada Gay que havia sido detectado no dia anterior. Com Jaccard os assuntos não sofreram união. Ambos se mantiveram de maneira independente no dia 07 de junho, o da campanha OBoticário sofrendo retração e o da Parada Gay passando por uma grande expansão. No entanto, os *tweets* que se tratavam da campanha OBoticário publicados no dia 07 de junho, refletiam comentários a respeito das fantasias com a embalagem de OBoticário que foram utilizadas na parada Gay[15] e brincavam com o surgimento da expressão "Raio Boticarizador" ou "Efeito Boticário", que fez piada mostrando participantes da Parada Gay e heterossexuais sendo transformadas após usarem produtos da marca.[16] Tal relação, deixa evidências de que no dia em que ocorreu a Parada Gay, a repercussão de OBoticário estaria estreitamente associada ao evento. Mais uma vez, a medida *Fading* que aproxima os assuntos (também) cronologicamente permitiu a detecção desta similaridade entre os dois *clusters*, unindo-os em apenas um.

No dia 08 de junho, o ***cluster*** referente a Parada Gay passou por uma **retração** e foi **detectado** novamente o ***cluster*** que se referia a pornografia Gay. Neste dia, o ***cluster*** referente

---

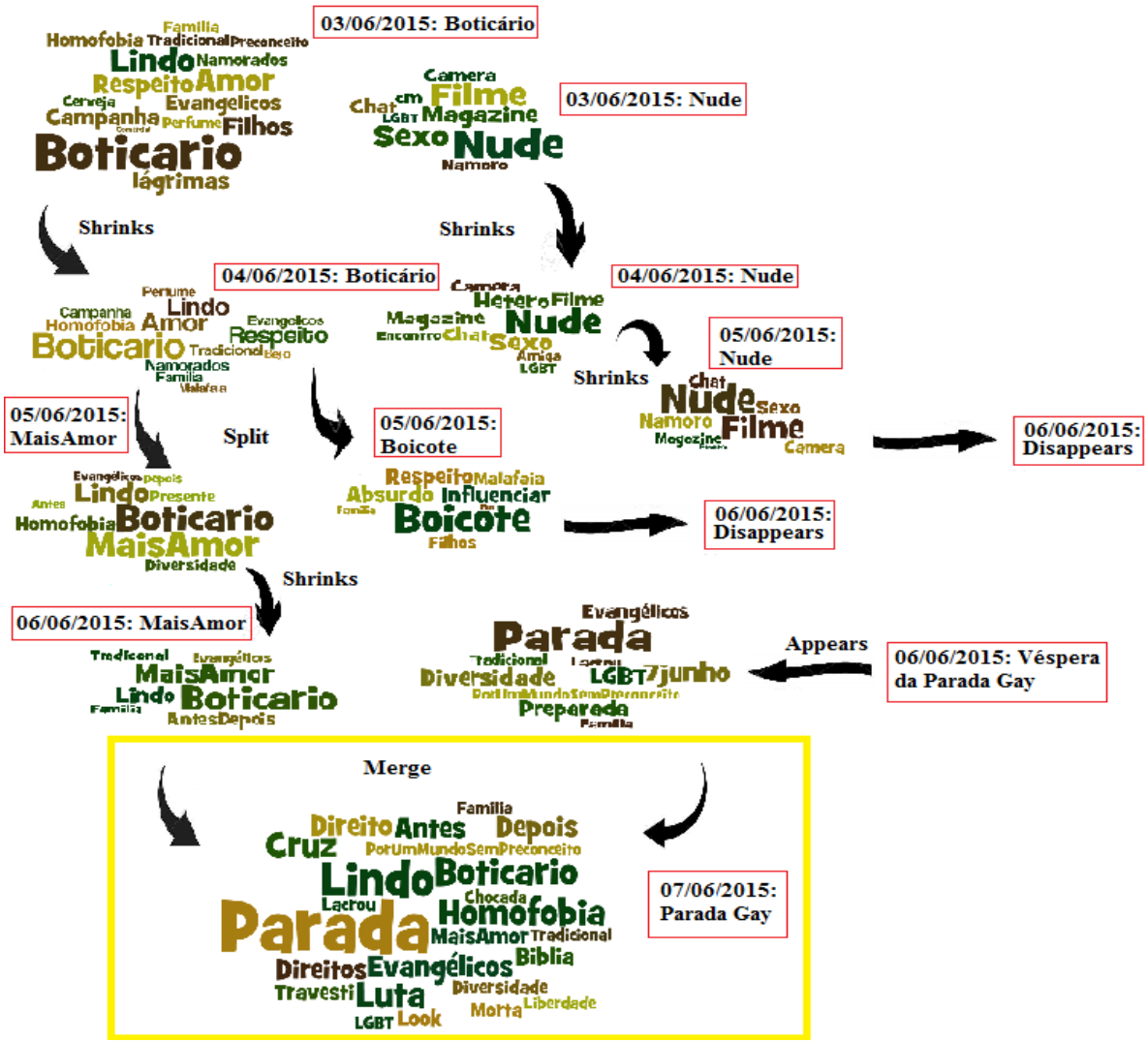
[15] http://noticias.uol.com.br/ultimas-noticias/agencia-estado/2015/06/07/namorados-se-fantasiam-de-embalagem-de-o-boticario-na-parada-gay.htm

[16] http://www.correio24horas.com.br/detalhe/noticia/veja-como-foi-o-efeito-boticario-na-parada-gay-de-sao-paulo/?cHash=0ced0a270cc4e061674af62403ccf5d2

a Parada Gay apresentou maior quantitativo de *tweets* que o mesmo *cluster* detectado no dia 08 de junho com Jaccard. Isso se explica pela união ocorrida no dia 07 de junho entre o tema Parada Gay e OBoticário que ampliou o volume de *tweets* deste *cluster*. Em Jaccard, no dia 08 de junho o ***cluster*** sobre OBoticário havia **desaparecido**, logo, muitos dos *tweets* que ainda haviam referentes a este tema no *snapshot* analisado, acabaram sendo eliminados como *outliers*.

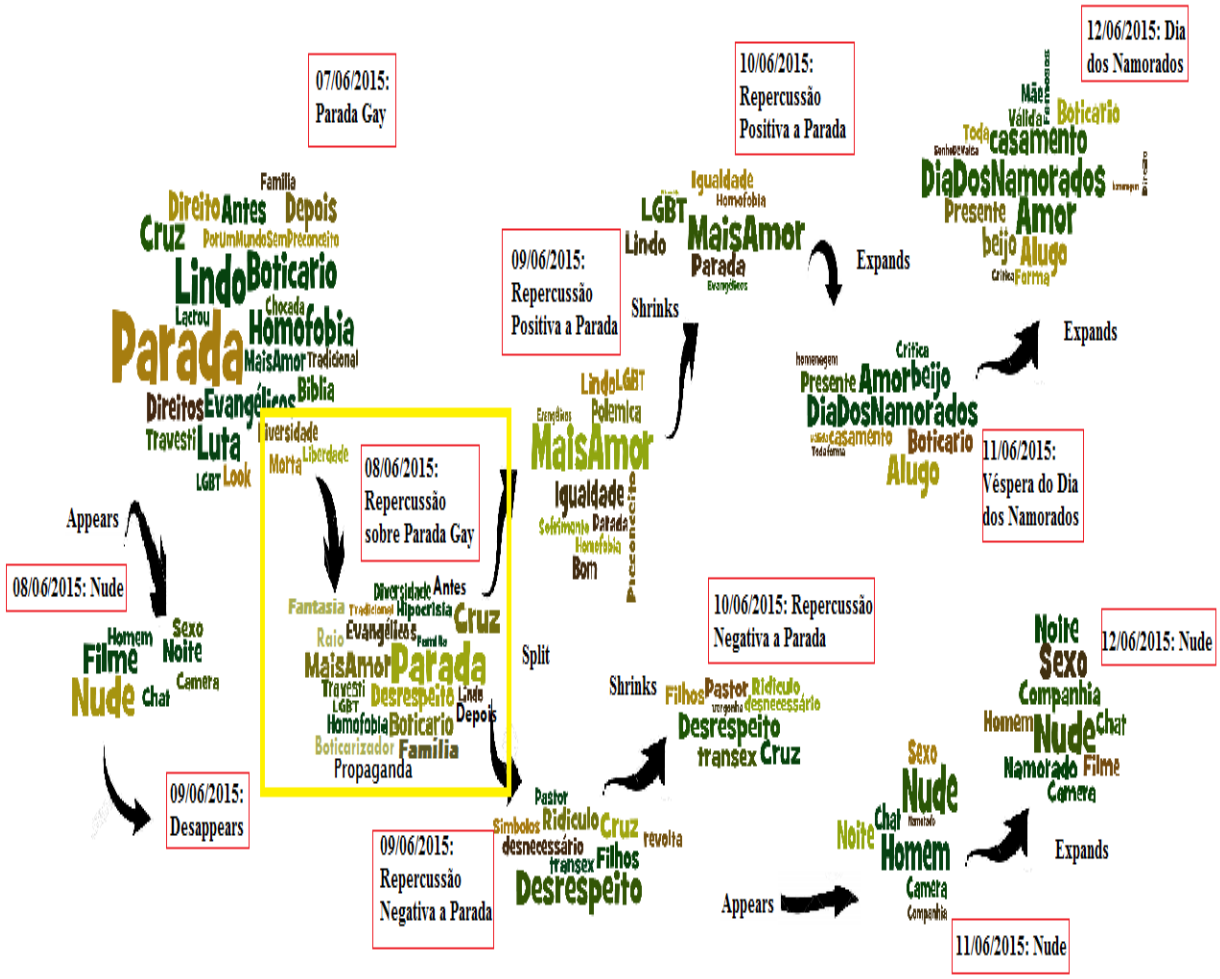
Dos dias 09 a 12 de junho, as transições apresentadas foram as mesmas detectadas na análise anterior com a medida de Jaccard. A seguir, nas Figuras 25 e 26, é apresentada por meio das nuvens de palavras, uma visão geral das evoluções sofridas por este assunto ao longo dos dias de análise.

Figura 25- Evolução dos assuntos sobre o tema "Gay" com Fading entre 03/06/15 e 07/06/15



Fonte: Elaborado pelo Autor.

Figura 26 - Evolução dos assuntos sobre o tema "Gay" com Fading entre 07/06/15 e 12/06/15



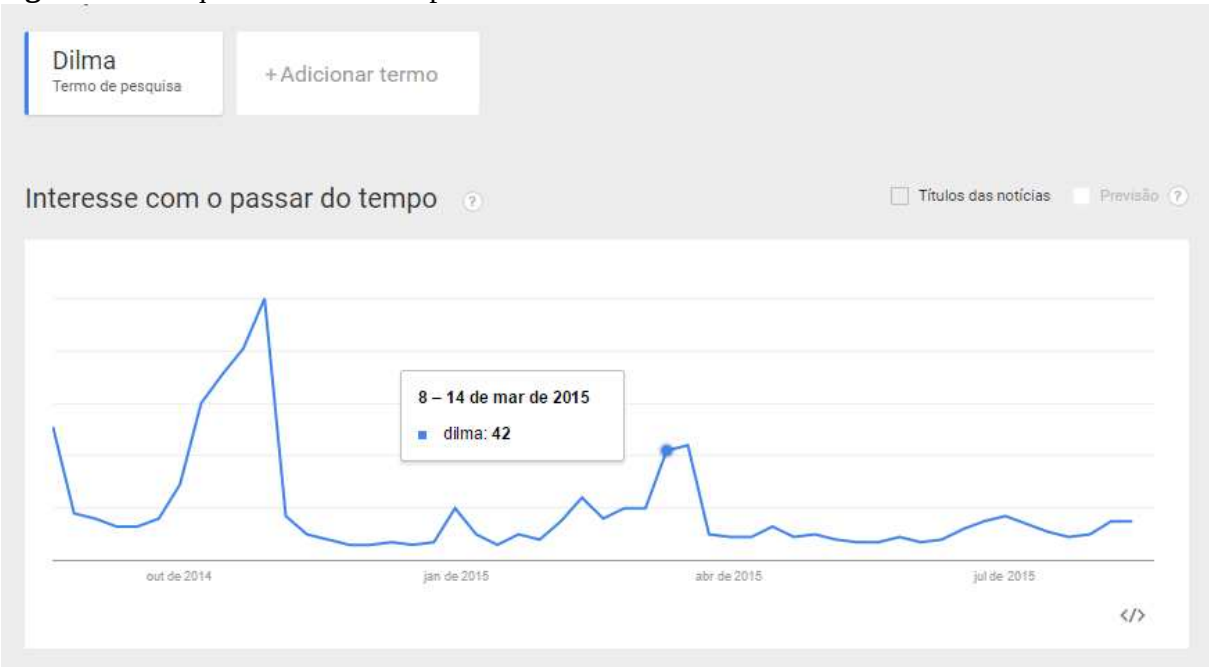
Fonte: Elaborado pelo Autor.

6.4 Validação dos resultados obtidos

O Google Trends[17] é uma ferramenta do Google que informa, por meio de gráficos, os termos mais buscados no site em um determinado período de tempo. O eixo horizontal do gráfico representa tempo e o vertical é a frequência com que um termo é procurado. A Figura 27 apresentada a seguir, contém um gráfico gerado pela ferramenta que mostra a frequência com que o termo “Dilma” foi buscado entre o período de outubro de 2014 e julho de 2015. O gráfico mostra que, após o período eleitoral que foi em outubro de 2014, o intervalo onde as buscas pelo termo “Dilma” ocorreram de forma mais intensa foi o período analisado neste trabalho. Entre 08 de março e 14 de março de 2015, as buscas pelo termo chegaram a atingir 42 pontos de frequência de um total de 100. Esses dados retratam a repercussão que este assunto alcançou no período da análise dos dados.

[17] https://www.google.com.br/trends/?hl=pt-BR

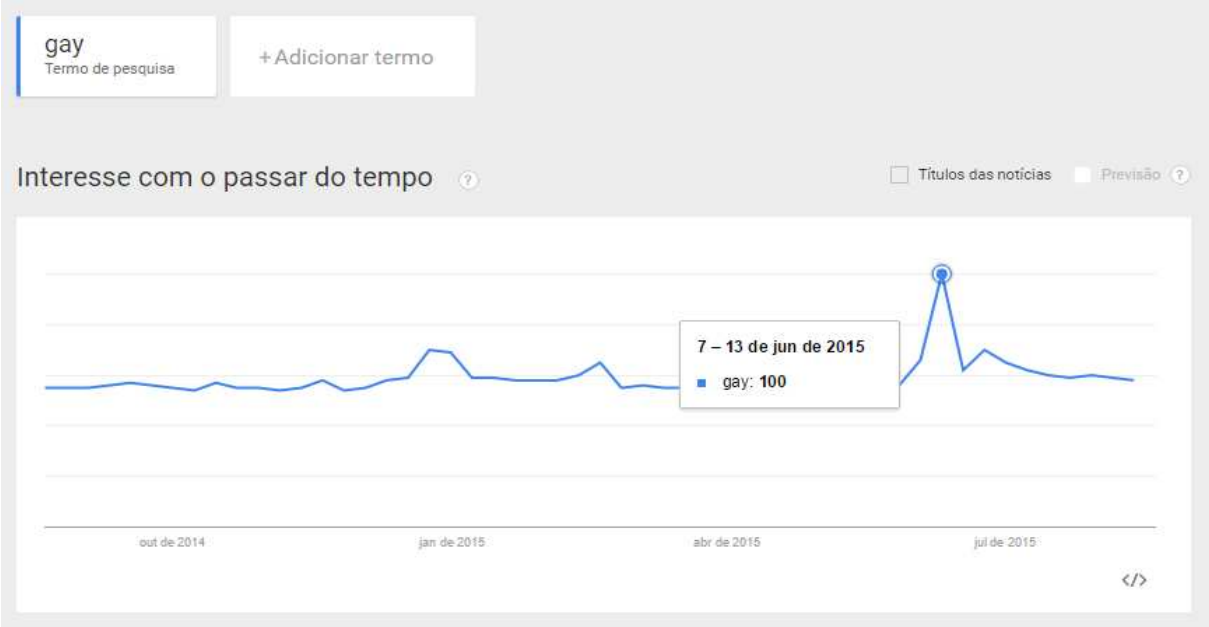
Figura 27- Frequência de buscas pelo termo "Dilma" entre 10/2014 e 07/2015



Fonte: Google Trends

O gráfico apresentado na Figura 28 mostra que a busca pelo termo “Gay” atingiu o pico de 100 na escala de frequência do Google Trends no período analisado neste trabalho. Este dado nos permite concluir que de fato, este foi um período no qual houve um intenso aumento nas pesquisas por esse termo em função dos acontecimentos que estavam ocorrendo.

Figura 28 - Frequência de buscas pelo termo "Gay" entre 10/2014 e 07/2015



Fonte: Google Trends

Essa ferramenta, embora útil para detectar a frequência da repercussão de temas na internet e demonstrar a expressividade dos assuntos analisados, não é capaz de detectar padrões de evolução dos assuntos, como por exemplo, a união e divisão dos mesmos apresentados neste trabalho.

Para além dos dados apresentados nos gráficos disponibilizados pela ferramenta Google Trends, diversos veículos de comunicação, como referenciados nas seções anteriores onde constam os resultados das análises, noticiaram os acontecimentos analisados neste trabalho demonstrando que o resultado obtido realmente estava de acordo com os acontecimentos reais.

Os resultados observados nas clusterizações com a medida Jaccard e com a medida *Fading* concordaram em sua grande maioria. No entanto, nos dias em que houve diferenças nos resultados pôde-se observar uma maior precisão na clusterização que utilizava a medida de similaridade *Fading*, visto que os resultados obtidos com esta medida de similaridade estavam de fato, mais coerentes com o ocorrido no período analisado (proximidade de assuntos sintaticamente e temporalmente). Este resultado ressalta a relevância que o atributo de tempo apresenta para uma maior precisão em análises de redes sociais, visto a probabilidade que postagens publicadas em um mesmo período de tempo têm de se referirem a assuntos similares.

7 CONSIDERAÇÕES FINAIS

Esse trabalho teve como finalidade demonstrar a eficácia do uso das técnicas de evolução de *cluster* como um recurso promissor para monitorar as transições de assuntos nas redes sociais. As etapas de coleta, pré-processamento, clusterização, evolução de *clusters* e análise dos resultados foram aplicadas em duas bases de dados do Twitter que se tratavam de assuntos distintos, e, em ambos os casos, foi possível observar de forma clara como os assuntos evoluíram ao longo dos dias, sendo factível detectar aparecimento, expansão, retração, divisão, união e o desaparecimento desses assuntos.

Além disso, na etapa de clusterização foram aplicadas, de maneira dissociada, duas medidas de similaridade nas bases de dados analisadas: Jaccard e *Fading*. Foram verificadas as semelhanças e diferenças nos resultados obtidos por meio de ambas, observando qual medida apresentou resultados mais similares aos que ocorreram durante a coleta de dados. Nesse contexto, por considerar o horário de publicação do *tweet* no cálculo da similaridade, possibilitando uma aproximação dos assuntos que estão perto cronologicamente, a Medida de Similaridade *Fading* apresentou resultados mais precisos que a medida Jaccard (de acordo com os acontecimentos do período da Coleta). Tais resultados, demonstrados ao longo da Seção 5, ressaltam que no contexto de Redes Sociais, o horário em que as publicações são feitas tem relevância para uma clusterização mais exata, tendo em vista a grande possibilidade de assuntos publicados num mesmo período de tempo se tratarem do mesmo contexto.

Vale ressaltar que, este tipo de aplicação é recente no contexto de redes sociais, no entanto apresenta-se bastante útil para usuários interessados em acompanhar a evolução de determinados acontecimentos ao longo do tempo, sobretudo, para aqueles usuários ou órgãos responsáveis por tomadas de decisão em relação ao assunto sobre o qual as evoluções estão sendo monitoradas. A partir dessa análise, as organizações podem se beneficiar dos resultados para os mais diversos fins, como elaborar estratégias de *marketing*, observar tendências do mercado, personalização de conteúdos e serviços dentre outros.

Como trabalhos futuros, o objetivo é propor uma nova medida de similaridade direcionada ao contexto de redes sociais. Uma medida que não avalie apenas a variável de tempo e similaridade sintática entre os textos, mas informações do contexto da publicação (localidade, informação do usuário que a postou...). Também é viável aplicar outras técnicas de Processamento de Linguagem Natural para melhor estruturar os textos que serão minerados, no intuito de se obter melhores resultados. É possível realizar um estudo mais profundo da variação dos parâmetros utilizados no DBSCAN para este contexto de

clusterização em redes sociais, bem como, experimentar diferentes variações do tamanho do snapshot. Neste trabalho foi utilizado um dia para cada snapshot, pode-se verificar como a estratégia se comporta para diferentes granularidades de tempo. Para os próximos assuntos a serem coletados, serão coletados também (durante o mesmo período) os *trending topics* do Twitter e do Google Trends. Isso ajudará a formar um ground truth para avaliar a qualidade da solução. Além disso, pode-se experimentar o processo seguido neste trabalho em outras redes sociais, como o Facebook, Whatsapp (histórico de conversas), dentre outras onde o emprego da evolução de *clusters* seja coerente na extração de informações relevantes.

REFERÊNCIAS

ARANHA, C.N. Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. 2007. **144 f. Tese (Doutorado em Engenharia Elétrica) – Pontífica Universidade Católica do Rio de Janeiro**, Rio de Janeiro. 2007

ESTER, Martin et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Kdd**. 1996. p. 226-231.

FARHI, Paul. The twitter explosion. **American Journalism Review**, v. 31, n. 3, p. 26-31, 2009

FIELDING, Roy. Representational state transfer. **Architectural Styles and the Design of Netowork-based Software Architecture**, p. 76-85, 2000.

JAIN, Anil K. Data clustering: 50 years beyond K-means. **Pattern recognition letters**, v. 31, n. 8, p. 651-666, 2010.

JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: a review. **ACM computing surveys (CSUR)**, v.31, n. 3, p. 264-323, 1999.

KAUR, S. et al. **Concept drift in unlabeled data stream**. Technical Report, University of Delhi, 2009

KIM, Min-Soo; HAN, Jiawei. A particle-and-density based evolutionary clustering method for dynamic networks. **Proceedings of the VLDB Endowment**, v. 2, n. 1, p. 622-633, 2009.

KASZNAR, Istvan Karoly; GONÇALVES, Bento Mario Lages.Técnicas de agrupamento Clustering.**Institutional Business Consultoria Internacional (IBCI)**, 2009.

KWAK, Haewoon et al. What is Twitter, a social network or a news media?. In: **Proceedings of the 19th international conference on World wide web**. ACM, 2010. p. 591-600.

LEE, Pei; LAKSHMANAN, Laks VS; MILIOS, Evangelos E. Incremental cluster evolution tracking from highly dynamic network data.In: **Data Engineering (ICDE), 2014 IEEE 30th International Conference on.** IEEE, 2014. p. 3-14.

OLIVEIRA, Márcia; GAMA, Joao. Understanding clusters evolution. In:**Workshop on Ubiquitous Data Mining.** 2010. P. 16-20.

RODRIGUES, Barbosa et al. Characterizing the effectiveness of twitter*hashtags*to detect and track online population sentiment. In: **CHI'12 Extended Abstracts on Human Factors in Computing Systems**. ACM, 2012. p. 2621-2626.

RUSSEL, Mathew A. **Mining the social web**: Data Mining Facebook, Twitter, LinkedIn,Google+, GitHub and More. 2 ed. Sebastopol: O'reilly Media, Inc., 2013.

SILVA, Tércio Jorge da; COELHO DA SILVA, Ticiana L.. Descoberta dos perfis dos alunos da Universidade Federal do Ceará via Mineração de Dados. 2014. 2 p. TCC (Graduação em Sistemas de Informação) - **Universidade Federal do Ceará**, Quixadá 2014.

SILVA et al. Discovering frequent mobility patterns on moving object data. In:**Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems**. ACM, 2014. p. 60-67.

SPILIOPOULOU, Myra et al. Monic: modeling and monitoring cluster transitions. In:**Proceedingsof the 12th ACM SIGKDD international conference on Knowledge discovery and data mining**. ACM, 2006. p. 706-711.

TANG, Lu-An et al. A framework of traveling companion discovery on trajectory data streams. **ACM Transactions on Intelligent Systems and Technology (TIST)**, v. 5, n. 1, p. 3, 2013.

YIN, Jie et al. Using social media to enhance emergency situation awareness. **IEEE Intelligent Systems**, v. 27, n. 6, p. 52-59, 2012.

YIP, Andy M.; DING, Chris; CHAN, Tony F. Dynamic cluster formation using level set methods.**Pattern Analysis and Machine Intelligence, IEEE Transactions on**, v. 28, n. 6, p. 877-889, 2006.