MODELANDO COMPORTAMENTO EM REDES SOCIAIS: UM NOVO ALCANCE PARA OS MÉTODOS QUANTITATIVOS NO ESTUDO DA DINÂMICA SOCIAL

André Carvalho Silveira[♣]

Resumo

Em uma realidade onde milhões de pessoas se comunicam, se expressam e interagem em redes sociais digitais, é possível repensar o alcance dos métodos quantitativos aplicados ao estudo da dinâmica social. Sobre uma nova base massivamente abrangente, dinâmica em tempo real, e digitalmente quantitativa, métodos clássicos podem gerar novos resultados significativos na identificação de padrões e tendências de comportamento. Esse artigo apresenta um fluxo de trabalho desde a obtenção de dados de uma rede social, conversão dos dados em vetores de características e agrupamento de alguns integrantes da rede com comportamento similar. Os integrantes foram comparados perante a escolha dos meios de transporte, e agrupados por clusters hierárquicos. O presente estudo de caso pretende contribuir para a avaliação de como essa classe de dados pode ser modelada por métodos quantitativos para gerar novos insights sobre a dinâmica social.

Palavras-chave: Modelagem computacional, Métodos quantitativos, Mineração de dados, Redes Sociais, Big data.

[♣]Mestrando em Análise e Modelagem de Sistemas Ambientais pelo IGC/UFMG.

1- INTRODUÇÃO

As redes sociais digitais

Redes sociais são plataformas digitais que utilizam a internet para permitir a identificação, expressão e relacionamento entre pessoas (Costa, 2005). O que antes parecia apenas mais um produto a explorar as potencialidades de conexão da rede mundial de computadores hoje se apresenta como um ambiente de atividade social acessado cotidianamente por milhões de pessoas ao redor do globo. Apesar da diversidade entre plataformas, contextos culturais e políticos, e acesso as TICs (Dijst, 2004), ainda assim, a frequência de uso e abrangência das comunicações encontradas nessas redes atualmente são tão expressivas quanto a quantidade de dados gerados a cada segundo pelas mesmas (Segaran, 2007). A disponibilidade, a relevância e o formato digital dos dados, bem como o dinamismo de auto-atualização da base, configura um novo alcance para os métodos quantitativos, dos mais básicos aos mais sofisticados. Se esses dados são gerados diariamente em um formato passível de quantificação, emerge assim um oceano de novas possibilidades na identificação de padrões e tendências, e consequentes descobertas no campo social seja para a ciência, gestão pública, ou para o mercado empresarial.

O número de sistemas de redes sociais presentes na internet varia rapidamente e, nos dias de hoje, varia de forma crescente (Costa, 2005). Gundecha& Liu (2012) apresentam as redes mais relevantes em termos do número de integrantes que as compõem. Cada uma dessas redes apresenta características próprias ao definir mecanismos com os quais cada integrante se identifica, se expressa e se relaciona. Tais características definem a dinâmica dessas redes e embora sua estrutura principal tenda a se manter, tais mecanismos são atualizados eventualmente, abrindo novas formas de se medir expressões e interações sociais. Dentre as redes sociais disponíveis, o Twitter foi a escolhida para servir como base de dados neste trabalho. Os principais mecanismos de comunicação do Twitter são descritos na próxima seção para subsidiar o entendimento de como os métodos quantitativos foram aplicados. A escolha do Twitter se dá pela facilidade operacional, principalmente em termos da Twitter API, por utilizar publicações objetivas em pequenas unidades com formato bem definido, por apresentar mecanismos claros que permitem quantificação direta (hashtags, menções, etc), e por ser uma rede muito utilizada em dispositivos móveis, o que é útil para o tema de transportes, tema esse escolhido para o estudo de caso deste trabalho.

Twitter

O Twitter é considerado um microblog, ou seja, um canal de comunicação onde pessoas comuns e profissionais de comunicação podem publicar pequenos textos de no máximo 140 caracteres, tal unidade de publicação no Twitter é chamada de tweet. Os principais elementos que um tweet pode conter são: texto pleno, links, hashtags e menções. As hashtags são keywords que levam como sufixo o caractere '#' para indicar que um termo específico foi usado. Dessa forma outros usuários podem usar o mesmo termo em suas publicações, contribuindo para o assunto pertinente a umahashtag específica. O símbolo '#' contribui para que o termo não se degenere em pequenas variações sintáticas, além de indicar intenção de uso do termo em questão. Dessa forma o uso de hashtags se torna um elemento claramente quantitativo na rede. Menções por sua vez, são referências a um integrante da rede através de um termo que o identifica. Cada integrante do Twitter se identifica por um termo único que usa o prefixo '@'. O símbolo '@', assim como o símbolo das hashtags, torna as menções claras de serem tratadas e quantificadas. Mais informações sobre os mecanismos da rede podem ser obtidas no website oficial do Twitter (TwitterLearn More, 2013). Os elementos principais descritos são suficientes para o entendimento da base de dados utilizada neste trabalho, bem como sobre como os métodos quantitativos foram aplicados.

Twitter-API

O presente trabalho objetiva modelar dados de redes sociais de forma massiva, sendo assim as principais interfaces do Twitter não são suficientes para obter o montante de amostras desejado, pois são orientadas ao uso pessoal da rede o que torna o consumo dos tweets lento demais para montar uma base de dados suficientemente abrangente. Nesse sentido, para se obter a base de dados deste trabalho foi utilizada a Twitter API (ApplicationProgramming Interface) (TwitterInc, 2013). A Twitter API é outra interface oferecida pela rede porém orientada a software, ou seja, o conteúdo da rede pode ser acessado por um artefato computacional e não diretamente por um humano. Graças a Twitter API é possível obter quantidades massivas de tweets de forma sistemática a partir de pesquisas parametrizadas. Para citar alguns parâmetros de pesquisa que podem ser usados para se obter dados do Twitter, temos: os tweets podem ser obtidos por integrante, por keyword, por geolocalização, entre outros. A figura 1 apresenta um esquema de como os dados foram amostrados da rede social através da Twitter API.

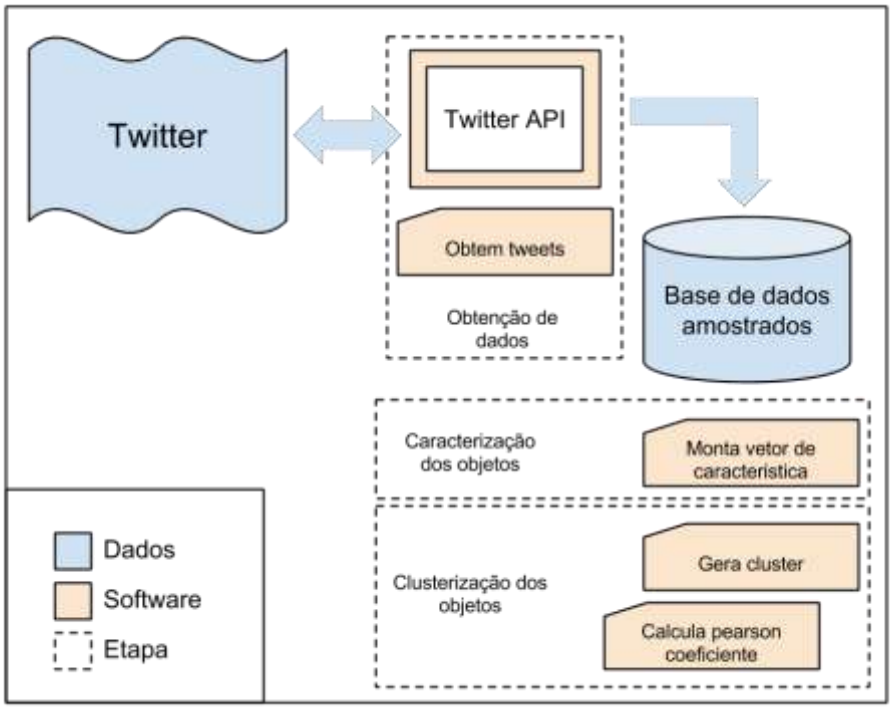

Figura 1: fluxo de dados e software desde a obtenção dos dados até a geração dos clusters.

2- METODOLOGIA

Base de dados

Através de pesquisas parametrizadas via Twitter API, tweets foram obtidos e armazenados em um banco de dados de amostra. O banco de dados utilizado possui estrutura simples com apenas duas tabelas: tb_tweets e tb_keywords. A tabela de tweets (tb_tweets) é o principal objeto a ser operado, sendo que além do texto de cada publicação (tweet) são colecionados metadados sobre os tweets como: autor da publicação, data e hora da publicação, parâmetros de pesquisa utilizados na Twitter API, data e hora da pesquisa de sua obtenção, entre outros. O banco foi populado com mais de 6000tweets, datados entre novembro de 2012 e novembro de 2013 distribuídos de forma heterogênea a priorizar tweets mais recentes. Os dados foram obtidos via Twitter API com pesquisas de mais de 50 keywords acerca do tema transporte. Tais pesquisas foram parametrizadas para obter tweets sobre as cidades de Belo Horizonte e São Paulo, embora ocorram outras localidades por não ser possível garantir a restrição espacial. Vale lembrar que toda informação que permita identificar os integrantes amostrados da rede foram removidas preservando o anonimato dos dados.

Artefatos computacionais

Para tornar possível a manipulação de dados em massa foram escritos algoritmos que permitem executar operações em lote de forma parametrizada. Os algoritmos foram implementados em software

usando a linguagem de programação Python (Richert&Coelho, 2013) e são aqui referenciados como scripts. Os scripts atuam desde a obtenção dos dados via Twitter API, o armazenamento em banco, a aplicação dos métodos quantitativos, até demais tratamentos secundários. A figura 1 apresenta como os principais scripts atuam em cada etapa do trabalho. Em linhas gerais as operações principais se dividem em três scripts: obtenção de tweets, montagem do vetor de características e geração dos clusters. Há também scripts para operações secundárias, como por exemplo o cálculo da métrica de distância usado para definir similaridade entre dois objetos no processo de geração dos clusters. A divisão do fluxo de trabalho em módulos é interessante para flexibilizar a metodologia. Com essa flexibilidade é possível, por exemplo, alterar a métrica de distância sem impactar as demais etapas do processo. Outra possibilidade é trabalhar com outra rede social usando outro script de obtenção de dados específico. A etapa de caracterização dos objetos é passível de vários testes até que se encontre características suficientes para o agrupamento desejado. Com a modularização do trabalho, os scripts responsáveis por essa etapa podem ser ajustados sem impactar as demais etapas.

Vetor de característica

Uma vez de posse de um banco de dados com tweets amostrados para diferentes integrantes da rede, o objetivo é caracterizar cada integrante com base no conteúdo de seus tweets, definir uma métrica para que se possa medir a similaridade a partir das características dos integrantes, e por fim agrupar integrantes com comportamento similar. Nesse caso o objeto modelado são integrantes da rede social, e o resultado esperado é um agrupamento de objetos onde integrantes que fazem parte do mesmo grupo possuem maior similaridade entre si se comparados a integrantes de outros grupos. Para que o agrupamento de integrantes seja efetivo, a forma como se caracteriza um integrante é crítica na metodologia. Sendo assim, para garantir um bom agrupamento é imprescindível que se defina um bom vetor de características.

Para caracterizar cada integrante é usado um artifício comum em técnicas de cluster e outros métodos quantitativos, os vetores de características. Esses vetores determinam quais características serão verificadas em cada integrante, em que ordem elas se dispõem, e qual a magnitude apresentada pelo integrante em cada característica. Cada característica tem sua magnitude compara entre dois integrantes, os quais são considerados tão similares quanto forem próximas suas magnitudes em cada característica. Como este trabalho tem interesse em análises quantitativas é útil que o vetor de características definido envolva variáveis numéricas. Uma forma simples e direta de caracterizar integrantes por suas publicações é avaliar a ocorrência de termos. Assim espera-se que integrantes que apresentam ocorrências próximas de determinados termos falem sobre determinado assunto com a mesma intensidade, ou tenham interesse equivalente por determinado assunto, sejam sensíveis a um

mesmo evento, ou sejam impactados de forma parecida perante um local ou objeto. Caracterizar integrantes por ocorrência de termos também apresenta a vantagem de considerar, sem tratamentos adicionais, a contabilização de hashtags e menções. Pois, Integrantes que usam a mesma hashtag serão caracterizados pela ocorrência da mesma em suas publicações, bem como integrantes que mencionam integrantes terceiros em comum também apresentarão tal similaridade na ocorrência de umamenção específica. Dessa forma a ocorrência de termos se mostra uma métrica razoável para caracterizar quantitativamente os objetos modelados no ensaio inicial.

Quebrando, tratando e contando palavras

Para utilizar a ocorrência de termos no vetor de característica é preciso extrair termos individuais de cada tweet de um integrante e contabilizar suas ocorrências. Para tal o script 'Monta vetor de característica' e complementares, recupera em banco de dados todos os tweets para um integrante específico, transforma cada tweet em uma lista de palavras individuais (termos) e conta a ocorrência de termos. O produto dessa etapa é um dicionário que contém todos os termos usados pelo integrante em suas publicações, bem como para cada termo o número de vezes que o mesmo ocorreu. Entretanto em primeira análise é possível perceber que não são todos os termos que trazem significado de forma contribuir para um vetor de característica eficaz. Preposições, artigos, e outras palavras que não agregam significado para caracterizar um integrante apresentam altos índices de ocorrência e assim prejudicam a montagem de um vetor eficaz. Outro defeito da contagem geral de termos são palavras muito raras e específicas o suficiente para ocorrer apenas uma vez, ou apenas nas publicações de um integrante. Esses termos não permitem comparações entre integrantes e assim apenas alongam o vetor sem contribuir para a caracterização dos objetos de forma satisfatória.

Os scripts da etapa de caracterização dos objetos, além de identificar e contabilizar termos, também removem termos muito ocorrentes e pouco significativos, e termos com ocorrência muito baixa. Outros tratamentos também são executados, como conversão do texto para caixa baixa e tratamento de caracteres especiais. Dessa forma, pequenas variações sintáticas dos termos não são consideradas, o que permite contabilizar de ocorrência dos termos de forma adequada.

Filtrandopalavras (Knowledge and Data driven approaches)

Após realizado o tratamento descrito ser aplicado aos dados amostrados têm-se uma lista de termos e suas ocorrências para cada integrante em análise. Cada termo listado contribui para caracterizar o comportamento do integrante na rede social em estudo. Algumas redes sociais são temáticas, ou seja, são criadas para discutir algum tema específico como, por exemplo, compras e vendas, oferta de emprego e relacionamento profissional, discussão e organização política, sugestões e comentários de músicas ou filmes, entre diversos outros temas centrais. No caso do Twitter, não existe

um assunto central, cada integrante publica sobre qualquer assunto de interesse e é comum que se encontre publicações sobre diferentes temas para um mesmo integrante (Gundecha& Liu, 2012). Nesse sentido, usar a ocorrência de todos os termos para caracterizar integrantes é uma abordagem generalista, sem foco em tema. Como se objetiva uma análise objetiva foi definido um tema para caracterizar os integrantes amostrados. Neste trabalho, foi definido com tema do estudo de caso: transportes. Assim deseja-se que o agrupamento em clusters dos integrantes retrate o comportamento em transportes, mais especificamente os meios de transportes utilizados pelos integrantes amostrados.

Para montar um vetor que caracterize os tipos de transporte utilizados pelo integrante é preciso filtrar os termos a serem considerados. Usuários do metro provavelmente apresentam uma convergência em termos relacionados a esse meio de transporte. Além de termos que se referem aas estações de metro e aos nomes das linhas, eventos que impactam usuários de metro podem ser comentados na rede social, como o caso de obras, alterações no serviço, e até as condições do transporte que hoje em dia são muito relevantes no cotidiano do cidadão, como lotações em horários de pico, greves e manifestações. Além disso, algumas hashtags são criadas e usadas coletivamente para tratar especificamente de assuntos pertinentes, e algumas vezes exclusivo, de usuários de metro. Menções também são termos que contribuem muito para caracterizar o uso de um meio de transporte. Como o Twitter possui integrantes que representam organizações é comum encontrar empresas que se identificam oficialmente na rede social, e assim são frequentes as comunicações entre usuários do serviço de transporte e empresas que oferecem serviços correspondentes. É possível notar altos índices de ocorrência de termos de menção a organizações relacionadas a determinado meio de transporte em tweets de seus usuários.

Para definir um elenco de termos eficaz para o vetor de característica é possível adotar basicamente abordagens data-driven ou knowledge-driven (Segaran, 2007). Caso a seleção dos termos relacionados a transportes seja orientada aos dados, a análise dos termos gerais contabilizados para todos os integrantes estudados definiria quais termos relativos a transporte devem permanecer e quais não contribuem com a comparação. Já a abordagem orientada a conhecimento, os termos são sugeridos a priori, antes de analisar os termos presentes no banco de amostra. A abordagem adotada nesse trabalho é knowledge-driven, sendo fixados 160 termos para compor o vetor de característica, sendo esses todos relacionados ao tema de transportes e divididos em classes metro, trem, ônibus, carro e bicicleta. Ou seja, cada um dos termos elencados faz relação a um ou mais meios de transporte urbano. Espera-se que um vetor de característica que registra a ocorrência desses 160 termos nas publicações de cada integrante, possa agrupá-los de forma a diferenciar integrantes que utilizam com mais frequência transportes diferentes. A tabela 1 apresenta alguns dos termos utilizados.

Tema	Termo	Classe
transporte	bike	bike
transporte	vazio	metro-trem-onibus
transporte	lento	onibus-carro
transporte	parado	onibus-carro
transporte	parados	onibus-carro
transporte	ponto	onibus
transporte	pontos	onibus
transporte	fluindo	onibus-carro
transporte	onibus	onibus
transporte	nibus	onibus
transporte	trem	trem
transporte	trens	trem
transporte	carro	carro
transporte	lenta	onibus-carro
transporte	lentamente	onibus-carro
transporte	flex	carro
transporte	biciclet	bike
transporte	pedal	bike
transporte	pedalar	bike
transporte	pedalando	bike
transporte	pedalei	bike

Tabela 1: Termos elencados para o tema transportes e classificados entre: carro, ônibus, trem, metro e bicicleta.

Matriz geral de comportamento de transporte

A tabela 2 apresenta a ocorrência de cada termo nas publicações de cada integrante. É possível notar como a ocorrência dos termos varia para cada integrante. Como cada termo está atrelado a um ou mais meios de transporte, a análise de ocorrência dos 160 termos pode identificar integrantes com o mesmo comportamento, ou seja, pessoas que utilizam com mais frequência os mesmos meios de transporte. É interessante notar que como cada termo escolhido para compor o vetor de característica não se associa apenas a um meio de transporte, o agrupamento final esperado não é crisp (Bezdek&Pal, 1998). Ou seja, o método agrupa integrantes com comportamento similar entre a escolha dos meios de transporte, mesmo para integrantes que utilizam mais de um meio, por exemplo, trabalham de metro e utilizam o carro para estudar após o trabalho.

	Termos											
Integrante	metr	velocidade	via	nibus	ciclofaixa	av	ciclovia	bike	linha	obras	trem	...
Integrante 1	2	1	1	6	2	5	18	58	2	1	4	...
Integrante 2	25	4	8	1	0	0	0	0	6	0	26	...
Integrante 3	2	0	2	0	0	0	0	0	0	0	1	...
Integrante 4	17	5	2	8	0	2	0	0	8	0	33	...
Integrante 5	1	1	1	7	0	0	0	0	4	0	51	...
Integrante 6	2	0	4	1	0	1	0	0	6	0	26	...
Integrante 7	0	0	2	0	0	2	0	1	5	1	13	...
Integrante 8	0	1	0	0	0	0	0	0	0	0	5	...
Integrante 9	22	0	7	0	0	0	1	2	6	4	2	...
Integrante 10	14	6	5	5	0	0	0	0	38	3	29	...
...	...	...	...	...	...	...	...	...	...	...	...	...

Tabela 2: Matriz geral de característica para 10 integrantes analisados. Apenas alguns dos 160 termos são apresentados.

Definição de clusters

Métrica de similaridade

Uma vez que todos os integrantes em análise foram caracterizados, é preciso definir uma métrica de similaridade, também chamada de métrica de distância, para comparar quantitativamente os integrantes. Comparando os integrantes uns aos outros é possível identificar clusters de similaridade. Como o vetor de característica foi criado a partir de termos relacionados ao tema de transportes, espera-se que a partir de um conjunto inicial de integrantes possa-se obter sub-grupos de integrantes que utilizam os mesmos meios de transporte. A métrica de similaridade adotada foi o coeficiente de pearson (Gundecha& Liu, 2012). Essa métrica foi escolhida por ser apta a trabalhar com vetores quantitativos, e por ser análoga a um espaço matemática de N dimensões no qual pontos se localizam próximos ou distantes dependendo de como se projetam em cada eixo. Nesse caso os eixos são os termos elencados nos vetor de característica, os pontos são os objetos modelados (integrantes da rede social), e as coordenadas são as ocorrências de cada termo para determinado integrante. Neste caso os integrantes estão sendo posicionados em um espaço de 160 dimensões.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Figura 2: Fórmula para calcular o coeficiente de Pearson. Fonte: adaptado de [7].

Um script foi escrito para calcular o coeficiente de pearson entre dois vetores de característica. Quanto mais próximo for o número de ocorrências de cada termo mais similares são considerados os integrantes comparados. Conforme ilustrado na figura 2, o cálculo do coeficiente de pearson é simples e não demanda muito tempo para ser processado (Coeficiente de Pearson, 2013). A figura 2 apresenta a fórmula básica de cálculo do coeficiente de pearson. Entretanto, uma nova métrica pode ser utilizada sem ser preciso atualizar as demais etapas do fluxo de trabalho, basta substituir o respectivo script por outro que segue os mesmos padrões de entrada e saída de dados.

Cluster hierárquico

Objetos caracterizados e comparáveis podem ser agrupados por diversos métodos quantitativos (Bezdek&Pal, 1998). Utilizaremos clusters para agrupar os integrantes analisados por comportamento de transporte. Métodos de identificação de clusters são úteis para visualizar agrupamentos inerentes em uma base de dados. É importante que o número de grupos a serem gerados não seja definido a priori, pois um agrupamento totalmente orientado a dados (data-driven) permite o reconhecimento de padrões existentes em uma massa de dados mas que não são claros em uma análise tabular simples. Com o intuito de manter a simplicidade deste ensaio foi escolhido o método de cluster hierárquico para agrupar os objetos modelados. Assim, os objetos (integrantes da rede social) são agrupados em iterações. Cada iteração identifica os objetos mais similares entre si e os agrupa em um novo objeto que será comparado aos demais na próxima iteração. Como cada iteração funde objetos similares em novos objetos, o processo termina com apenas um objeto resultado da fusão de todos os iniciais. Quando dois objetos são fundidos, o novo objeto criado possui como valores de seu vetor de característica a média aritmética dos valores dos vetores dos objetos em fusão. Assim cada característica do novo objeto possui a magnitude média da mesma característica dos objetos fundidos. Após o processo de clusterização hierárquica é possível ver o resultado de cada iteração e assim observar como os objetos se organizam em relação a similaridade.

Análise em dendograma

Uma forma de visualização de dados útil para avaliar o resultado dos clusters hierárquicos é o dendograma. O dendograma é um diagrama que exibe a relação entre vários objetos ordenadamente. Através de um dendograma é possível notar com clareza quais objetos se apresentam mais próximos na primeira iteração, bem como reconhecer quantas iterações foram necessárias para dois objetos específicos se agruparem. Quanto mais iterações são necessárias para que dois objetos se agrupem, menos similares esses dois objetos são. O resultado da clusterização dos integrantes caracterizados por comportamento de transporte é apresentado no dendograma da figura 3. A figura 3 ilustra quais integrantes foram agrupados já nas primeiras iterações e quais foram reunidos apenas no fim do

processo. Para complementar o entendimento de similaridade por vetor de característica as figuras de 4 a 11 apresentam gráficos que permitem visualizar quais meios de transporte são mais referenciados por cada integrante em análise. A série de gráficos da figura 4 a figura 11 retratam o agrupamento por clusters hierárquico apresentada na figura 3. É possível notar como os integrantes que foram inicialmente agrupados se distribuem pelos meios de transporte de forma mais similar do que os integrantes agrupados posteriormente.

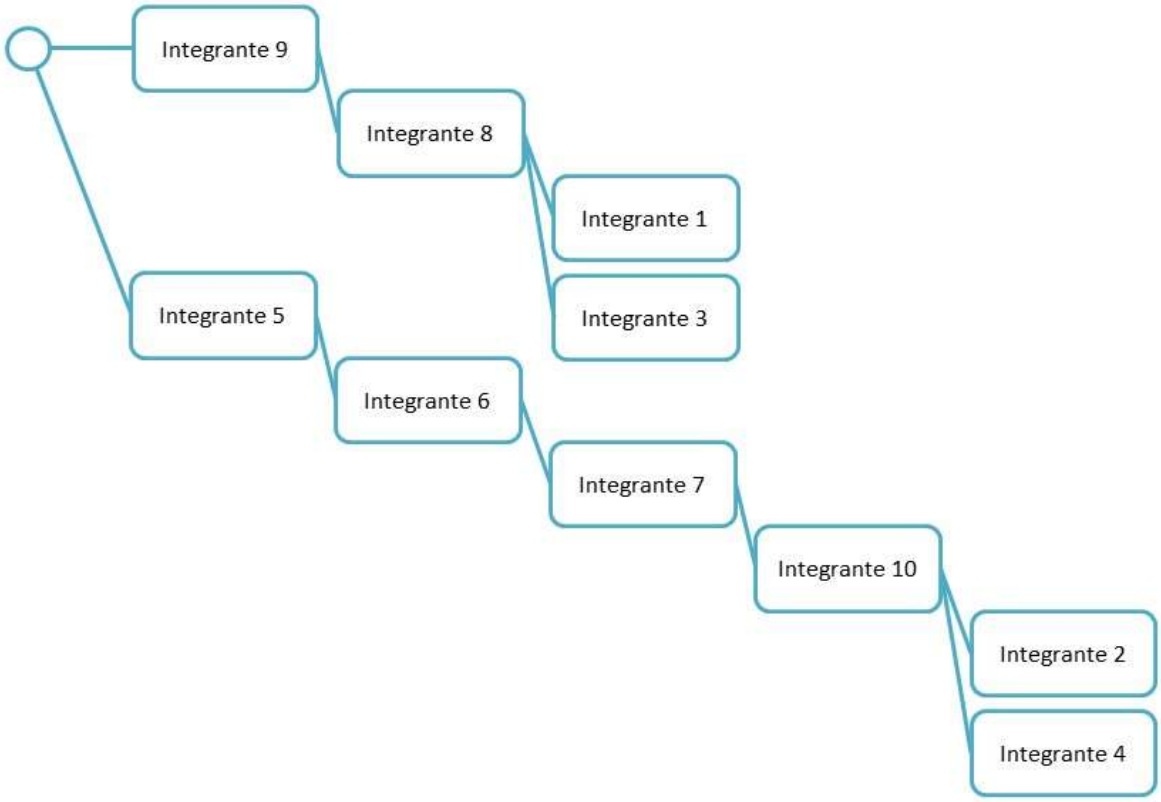

Figura 3: Dendograma de clusters hierárquicos para alguns integrantes analisados.

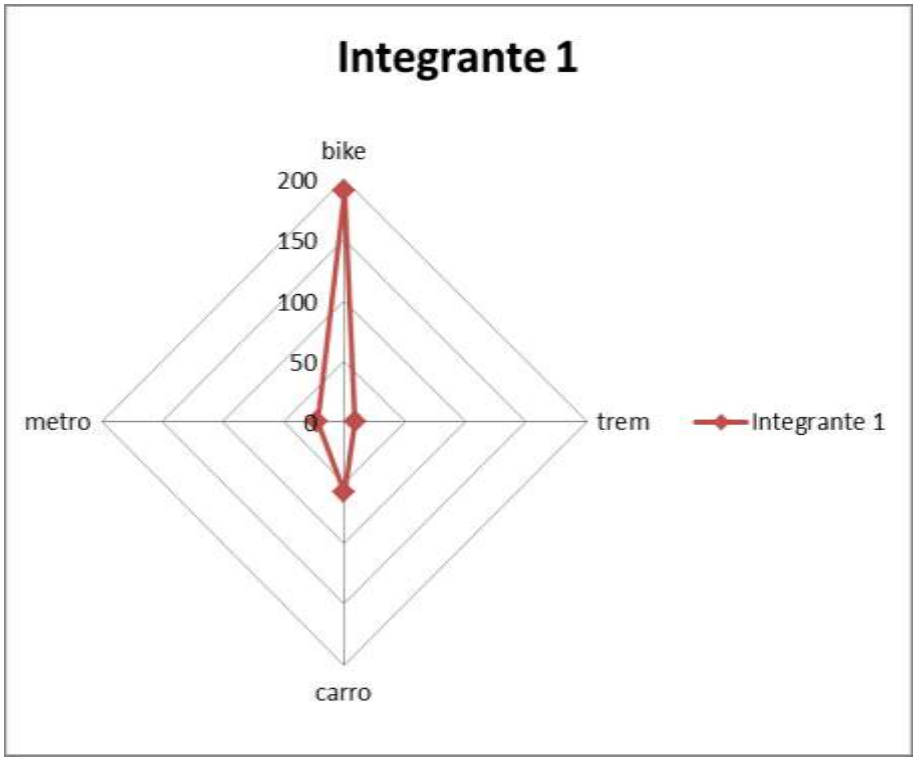

Figura 4: Comportamento do integrante 1 em relação aos meios de transporte referenciados em suas publicações. Nota-se a tendência de referenciar mais a bicicleta com meio de transporte em relação aos demais meios em estudo.

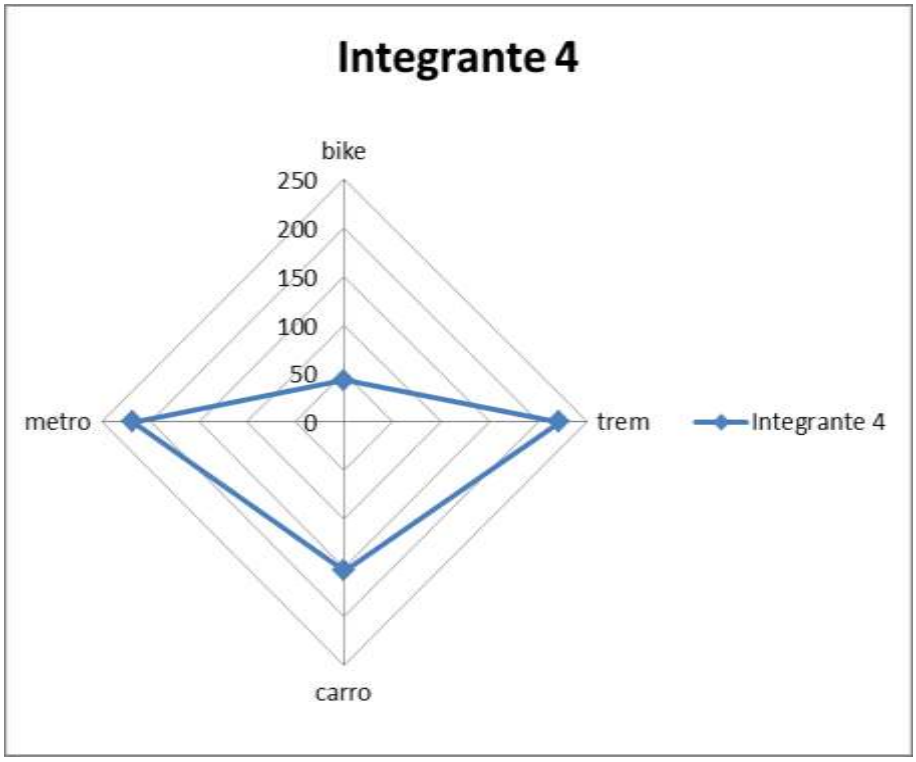

Figura 5: Comportamento do integrante 4 em relação aos meios de transporte referenciados em suas publicações. Nota-se maior número de referências a metro e trem, número de referências médio a carro e pouca referência a bicicleta.

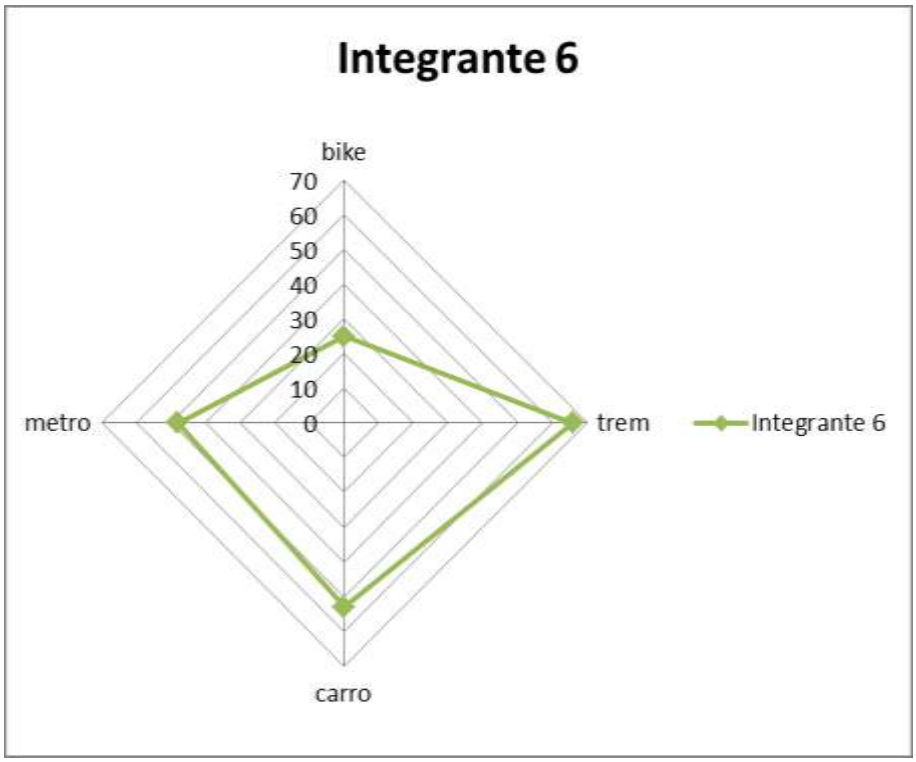

Figura 6: Comportamento do integrante 6 em relação aos meios de transporte referenciados em suas publicações. Notam-se mais referências a trem, número de referências médio a carro e metro, e menos referências a bicicleta.

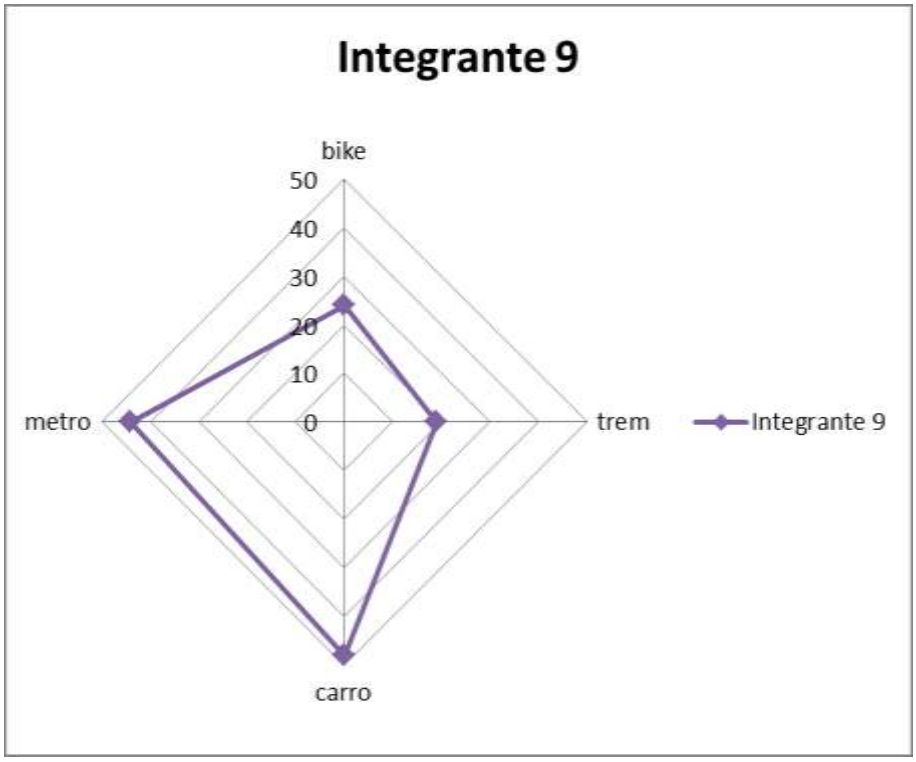

Figura 7: Comportamento do integrante 9 em relação aos meios de transporte referenciados em suas publicações. Nota-se tendência as referências de carro e metro, e menores números de referências a trem e bicicleta.

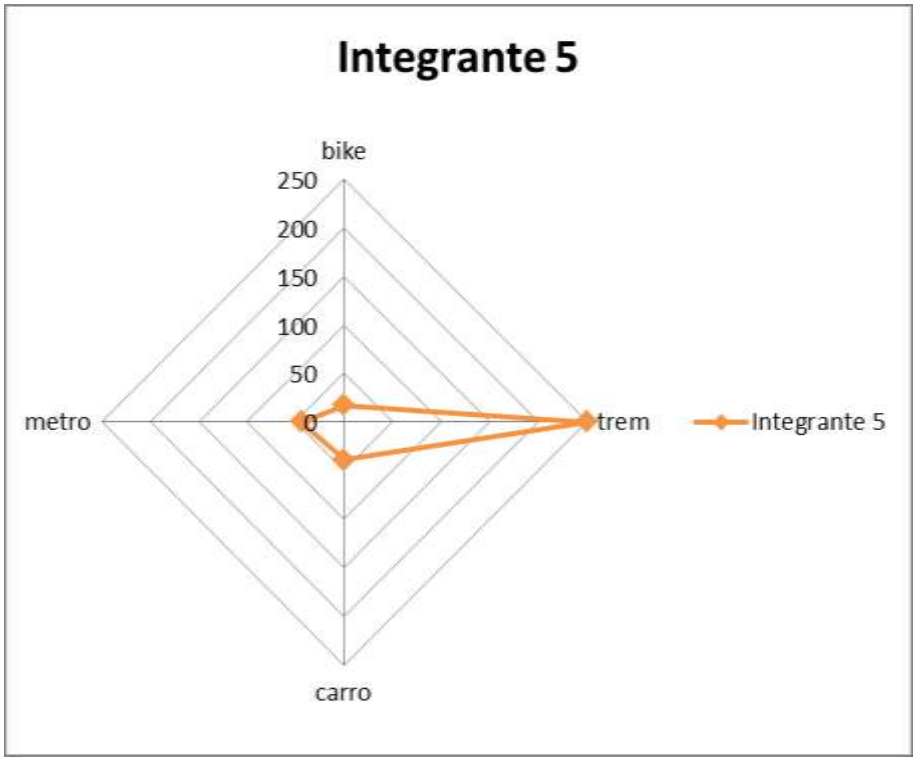

Figura 8: Comportamento do integrante 5 em relação aos meios de transporte referenciados em suas publicações. Nota-se clara tendência ao trem, e baixo número de referências a carro, metro, e bicicleta.

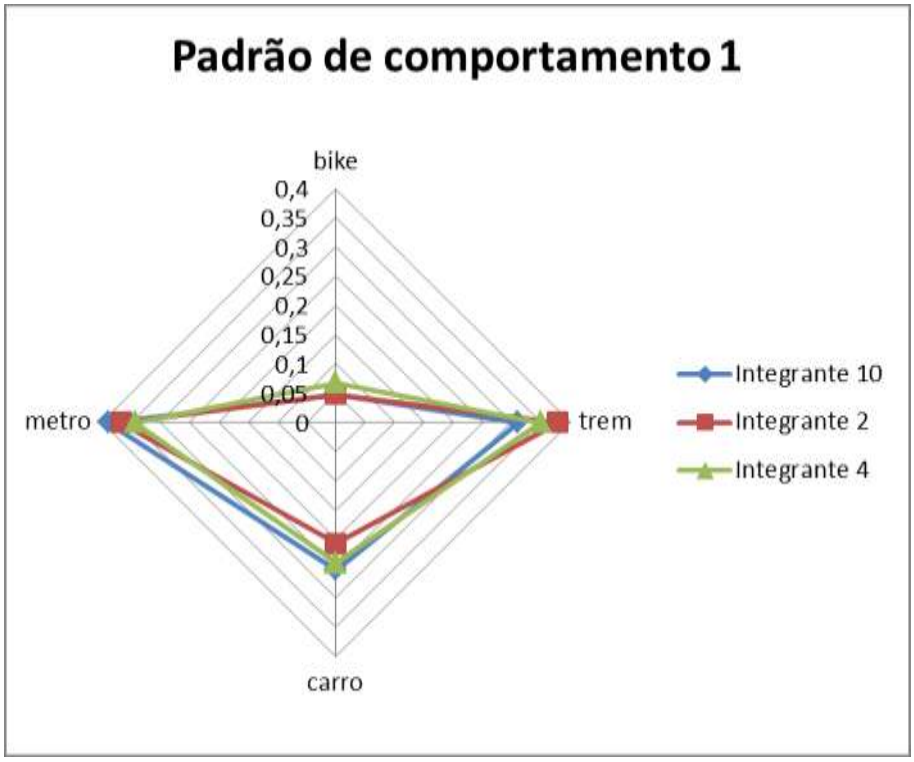

Figura 9: Padrão 1 de comportamento entre integrantes. Os meios de transportes estudados são referenciados de forma proporcional entre os integrantes desse grupo. O padrão infere similaridade entre os integrantes.

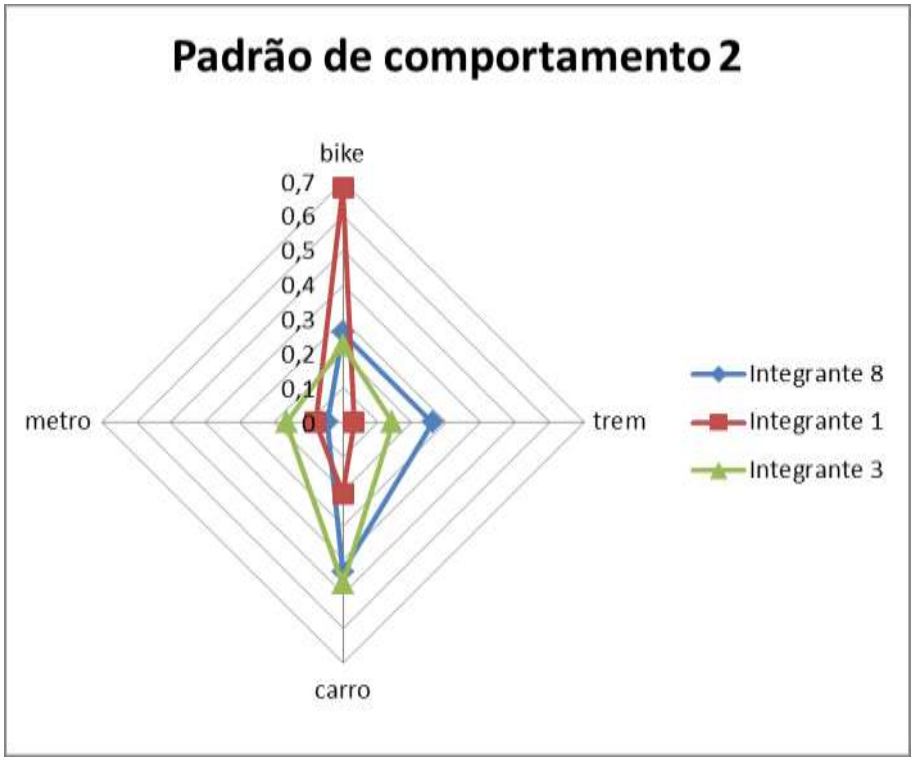

Figura 10: Padrão 2 de comportamento entre integrantes. Os meios de transportes estudados são referenciados de forma proporcional entre os integrantes desse grupo. O padrão 2 apresenta maior variância interna se comparado ao padrão 1, mas ainda assim infere similaridade entre os integrantes.

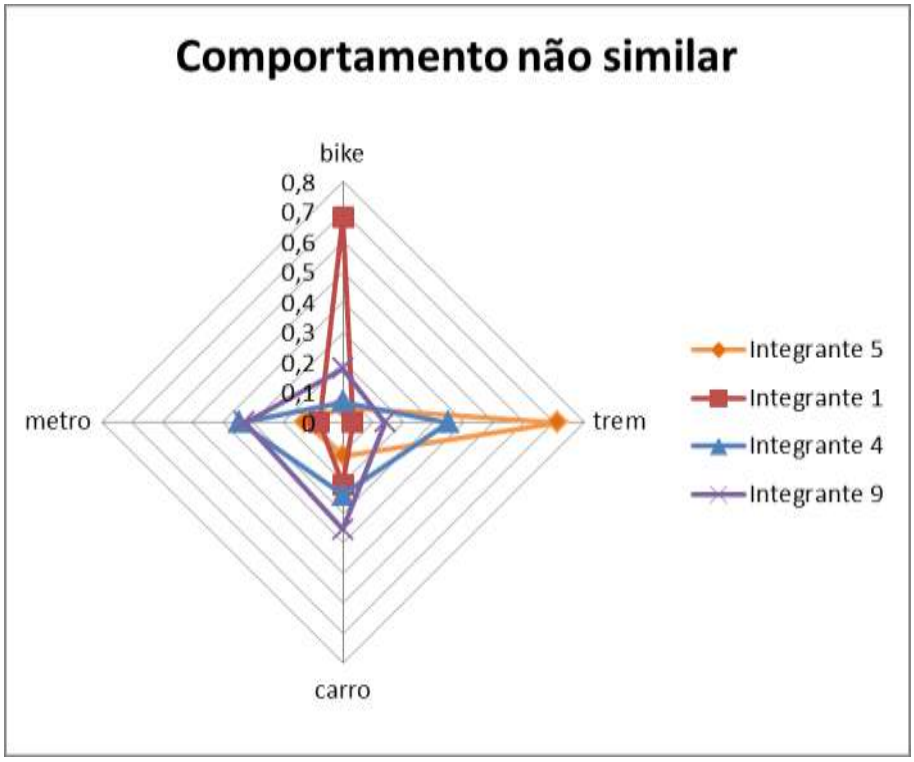

Figura 11: Integrantes com diferentes comportamentos em relação a referência aos meios de transporte. Os integrantes ilustrados apresentam baixa similaridade entre si.

RESULTADOS

Clusters por comportamento de transporte

A metodologia apresentada se mostrou apta para identificar pessoas que se referem de forma similar aos meios de transporte urbano. Os dados analisados não são gerados com rigor a representação da realidade. Na verdade, são dados gerados e publicados pelos próprios integrantes analisados no trabalho. Apesar da geração dos dados utilizados ser espontânea, não profissional e não sistemática, ainda assim espera-se que tais dados reflitam um aspecto da realidade de seu autor. Apesar da incerteza, é coerente entender os clusters resultados como indícios de comportamento de transporte. Devido a natureza dos dados, os resultados do trabalho devem ser usados de maneira adequada. Os resultados apontados não são suficientes para identificar o meio de transporte mais utilizado por determinado integrante da rede, entretanto são úteis para indicar qual meio de transporte é substancialmente relevante para um integrante específico. Essa informação não é tão objetiva e precisa se comparada a análises que usam como insumo dados levantados cientificamente. Por outro lado, são insights reais sobre a dinâmica social. Se considerada a abrangência desses dados, sua disponibilidade e atualização, somadas a inúmeros métodos quantitativos que podem ser combinados para extrair informações cada

vez mais assertivas, a mineração de dados sobre o big data se mostra como uma forma valiosa de estudos sociais.

A criticidade do vetor de características

É importante notar que qualidade da identificação de comportamentos similares através dos clusters é fortemente determinada pela escolha do vetor de característica. Se as características dos objetos não os diferem, ou mesmo se não os equivalem em nenhum ponto, logo a comparação entre os mesmos se torna pouco eficaz. É interessante que o vetor de característica seja sensível a propriedades peculiares dos objetos, mas que também permitam uma convergência entre os mesmos. Por essa razão, preposições, artigos e outras palavras muito recorrentes foram removidas do vetor, assim como termos raros ou muito específicos para serem citados por mais de um integrante. Outro ponto crítico sobre o vetor de características é o fato do mesmo ser responsável pela perspectiva da análise. No caso de modelagem de comportamento, o tema estudado será induzido pelo vetor. Neste trabalho o tema de transporte foi aplicado pela escolha de termos correlatos para compor as 160 ordenadas do vetor. Se este trabalho for reaplicado com apenas alterando o vetor para que passe a conter termos relacionados a consumo, os grupos de similaridade resultantes poderiam ser compostos de forma bem diferente dos resultados obtidos aqui.

Aplicabilidade da metodologia

Por fim, vale lembrar que apesar da simplicidade dos métodos apresentados, a metodologia descrita é toda automática. A partir de uma lista de integrantes da rede social em estudo, é possível obter dados para cada integrante listado, caracterizá-los perante um tema específico, e identificar comportamentos similares entre eles. A metodologia pode ser aplicada a tantos integrantes mais quanto houver disponibilidade da rede social fonte. É fato que um número muito grande de integrantes em uma única análise pode exigir um alto poder de processamento já que cada integrante precisa ser comparado com todos os outros. A mesma metodologia e infraestrutura discutidas no trabalho podem ser aplicadas em outro ensaio onde o número de integrantes se mantém, aumentando o número de tweets por integrante. Dessa forma cada objeto seria caracterizado com mais riqueza o que poderia resultar em um melhor fitness na modelagem de comportamento. A re-aplicação deste trabalho restringindo tweets por janelas de tempo poderia acompanhar a variação de comportamento dos integrantes estudados. Enfim, este trabalho pretende contribuir com as novas potencialidades dos métodos quantitativos aplicados a dinâmica social perante o big data, do que de fato com o avanço de métodos específicos.

Flexibilidade da metodologia

Alterando apenas o vetor de características a mesma metodologia é capaz de identificar similaridade de comportamento na ótica de outro tema. Também sem alterações na metodologia, é possível testar outra métrica de similaridade, além do coeficiente de pearson, ou substituir o método de clusters hierárquicos por outro método de agrupamento. Ainda com a mesma metodologia é possível substituir o script de obtenção dos dados e modelar insumos de outra rede social, desde que a mesma disponibilize uma interface API e se tenha um script respectivo com o mesmo padrão de fluxo de dados. Tal flexibilidade metodológica foi alcançada pela divisão do fluxo de trabalho em módulos.

CONCLUSÃO E DISCUSSÃO

Dados de mídias sociais, O que é e o que se diz que é

Dados com fonte em mídias sociais já foram considerados inadequados para estudos científicos por seu caráter não profissional. Esses dados são gerados de forma não sistemática e não assumem compromisso em relatar fatos da realidade. Entretanto hoje o chamado big data é utilizado em trabalhos científicos que exploram suas possibilidades (Gundecha& Liu, 2012). A abrangência apresentada por tais dados é sem precedentes, ainda mais sendo dados digitais, ou seja, já são criados em formato manipulável (Gundecha& Liu, 2012). Uma fonte de dados tão vasta somada as atuais potencialidades de poder computacional, confere aos métodos quantitativos uma nova e ampla gama de aplicações que podem contribuir em diversas áreas de pesquisa. Assim cabe o uso com cautela dessa classe de dados. Publicações em redes sociais não podem ser lidas como relatos fiéis a realidade, e sim como expressões individuais livres de qualquer rigor (Costa, 2005). Dessa forma as publicações de determinado indivíduo não relatam fatos, e sim expressões sobre os fatos. Respeitando a natureza do dado durante todo o processo metodológico é possível obter informações úteis para complementar inferências ou verificar hipóteses.

Nem todos são passíveis de modelagem

No decorrer do trabalho é possível notar que nem todos os integrantes de uma rede social podem ser modelados para um tema específico. No caso de comportamento em transportes, nem todos integrantes publicam a respeito de seus meios de transporte. Como não há um tema central proposto para discussão, cada integrante se expressa sobre assuntos, eventos, fatos, que lhes convém.

Por outro lado, nos dias de hoje, o uso das redes sociais é tão comum, que o número de integrantes é grande o suficiente para que seja possível obter amostras razoáveis nos mais variados temas. São necessários trabalhos que verifiquem a representatividade dessas amostras, uma vez que no escopo de uma cidade possivelmente os integrantes das redes sociais apresentam viés seja de idade, poder aquisitivo, grau de instrução, ou outros, apesar da popularização das Tecnologias de Informação e Comunicação, e da crescente inclusão digital (Dijst, 2004).

Novos dados, novas possibilidades

Em suma, o presente trabalho tenta contribuir o estudo sobre os novos limites dos métodos quantitativos a partir de uma nova base de dados que registra o comportamento milhões de pessoas, é disponibilizada voluntariamente, atualizada de forma coletiva praticamente em tempo real, e criada originalmente em formato digital. Diante desse recente cenário de fonte de dados, faz sentido ponderar rigor e abrangência para decidir entre o melhor uso do big data, e quem possivelmente converter um cenário atual onde se geram grandes volumes de gigabytes por hora em um cenário futuro onde se esses novos dados se traduzam em novos conhecimentos para a ciência de forma a permitir novas tecnologias seja para a gestão pública ou mercado de trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

Coeficiente de Pearson. (2013). Acesso em Novembro de 2013, disponível em Wikipedia: http://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson

TwitterLearn More. (2013). Acesso em Novembro de 2013, disponível em Twitter.com: https://discover.twitter.com/learn-more

Bezdek, J., &Pal, N. (Junho de 1998). Some new indexes of cluster validity. TransactionsonCybernetics IEEE, pp. 301-315.

Costa, R. d. (2005). Por um novo conceito de comunidade: redes sociais, comunidades pessoais, inteligência coletiva. Interface .

Dijst, M. (2004). ICTs and Accessibility: An Action Space Perspective on the Impact of New Information and Communication Technologies. In: Transport Developments and Innovations in an Evolving World. Berlim: Springer Berlin Heidelberg.

Gundecha, P., & Liu, H. (2012). Mining Social Media: A Brief Introduction. Tutorials in Operational Research Informs.

Richert, W., & Coelho, L. P. (2013). Building Machine Learning Systems with Python. Packt.

Segaran, T. (2007). Programming Collective Intelligence: Building Smart Web 2.0 Applications. O'Reilly.

Twitter Inc. (2013). Twitter for Developers. Acesso em Novembro de 2013, disponível em Twitter.com: https://dev.twitter.com/