UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

RICARDO CHAGAS RAPACKI

KANDOR – Um método de *clustering* para Análise de Conhecimento de Dinâmicas em Vizinhanças e Relacionamentos *Online* em Mapeamento Urbano

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Profa. Dra. Renata Galante

Porto Alegre
2018

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Rapacki, Ricardo Chagas

KANDOR – Um método de *clustering* para Análise de Conhecimento de Dinâmicas em Vizinhanças e Relacionamentos *Online* em Mapeamento Urbano / Ricardo Chagas Rapacki. – 2018.

15 f.:il.

Orientador: Renata Galante.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2018.

1.Cidades inteligentes. 2.Redes Sociais. 3.Planejamento Urbano. 4.*Clustering* espectral. I. Galante, Renata. II. KANDOR – Um método de *clustering* para Análise de Conhecimento de Dinâmicas em Vizinhanças e Relacionamentos *Online* em Mapeamento Urbano.

RESUMO

Com o surgimento de *smartphones* e redes sociais baseadas em localização (LBSNs), uma vasta quantidade de dados gerados por usuários se tornou disponível para análise em diversas áreas. Uma destas é a de planejamento urbano, que utilizando estas novas informações pode aprimorar e modernizar sua compreensão de dinâmicas de cidades e a interação entre seus locais e cidadãos. O objetivo deste trabalho é propor o método KANDOR para testar novas formas de representação de *cluster*s a partir de dimensões diferentes de dados de redes sociais, de modo a descobrir informações sobre dinâmicas de cidades e caracterização urbana. Características como avaliação de estabelecimentos, entropia (variedade de usuários) e popularidade podem revelar informações novas e um entendimento mais completo da estrutura e dinâmica da cidade e sua constante transformação. Enquanto os principais trabalhos relacionados utilizam um número reduzido de dimensões dos dados, o método proposto por esta dissertação visa se beneficiar da riqueza de informações existente em redes sociais. Uma pesquisa com usuários foi executada usando uma base de dados de *check-ins* de Porto Alegre, Brasil e aplicando *clustering* espectral com os modelos propostos. Os experimentos demonstraram que o uso das dimensões propostas contribui para gerar regiões com características diferentes como tamanho, coesão, desempenho e propagação. O método permite agregar uma vasta diversidade de informações de redes sociais para gerar visões diferentes e complementares da cidade. Logo, aplicando esses métodos em ambientes urbanos, governos e cidadãos podem compreender melhor suas cidades e construir cidades cada vez mais inteligentes.

Palavras-chave: Cidades inteligentes. Redes sociais. Planejamento urbano. *Clustering* espectral. Pesquisa quantitativa.

KANDOR – A Clustering Method for Knowledge Analysis of Neighborhood Dynamics and Online Relationships in Urban Mapping

ABSTRACT

With the emergence of smartphones and location-based social networks (LBSNs), a large amount of user-generated data has become available for analys in several areas. One example is urban planning, which with this new information can enhance and modernize the understanding of city dynamics and the interactions between venues and residents. The goal of this work is to propose the method KANDOR to generate new forms of cluster representation from different dimensions of social network data, so that information can be discovered about city dynamics and urban characterization. Characteristics such as venue rating, entropy (user variety) and popularity can reveal new information and a better understanding of the structure and dynamic of the city and its constant transformation. While the main related works use a reduced number of data dimensions, the proposed method by this dissertation aims to benefit from the extent of the information existing in social platforms. A user research is executed using a check-in dataset from Porto Alegre, Brazil and applying spectral clustering with the proposed models. The experiments demonstrated that the use of the proposed dimensions contributed to generate regions with different characteristics like size, cohesion, performance and propagation. KANDOR allows the aggregation of a large diversity of information from different social networks to generate different and complementary visualizations of the city. Hence, by applying these methods to urban environments, governments and citizens can better understand and build better sustainable cities together.

Keywords: Smart cities. Social Networks. Urban Planning. Spectral clustering. Quantitative research.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANOVA	Análise de Variância
API	*Application Programming Interface*
AV	Avaliação
HP	Horários de Pico
KANDOR	*Knowledge Analysis of Neighborhood Dynamics and Online Relationship*
LBSN	*Location-Based Social Networks*
PP	Popularidade
PRE	Precisão
PRO	Propagação
SDK	*Software Development Kit*
SOM	*Self-Organizing Maps*
SS	Similaridade Social
VU	Variedade de Usuários

Sumário

1 INTRODUÇÃO

Grandes volumes de dados gerados por usuário se tornaram disponíveis como fonte de informação quando *smartphones* e redes sociais se tornaram comuns no cotidiano das pessoas, permitindo novas descobertas sobre atividades e rotinas de pessoas (Noulas et al., 2011). Alguns exemplos entre fontes de informação disponíveis são redes sociais baseadas em localização (LBSN) como Foursquare[1], redes sociais de compartilhamento de mídia como Instagram[2] e plataformas de propósito geral como Facebook [3]e Twitter[4]. Os desafios que são encontrados nessa área são em geral os mesmos na área de *Big Data*, como integração de dados a partir de múltiplos fornecedores, gerenciamento de grandes volumes de dados e limpeza de dados.

Quando estas informações são utilizadas em áreas como planejamento urbano, redes sociais podem contribuir para a compreensão da cidade e suas regiões por uma fração do custo de métodos tradicionais e sistemas de sensores (Silva et al., 2012). Portanto, estas plataformas ajudaram a revolucionar a área e evoluir o conceito de *Smart Cities*, que define cidades com infraestruturas e serviços desenvolvidos e guiados por cidadãos e sensores (Frias-Martinez e Frias-Martinez, 2014).

Entretanto, implementar um processo automatizado de descoberta de conhecimento em cidades utilizando redes sociais não é uma tarefa fácil. Primeiramente, existem inúmeras plataformas sociais com diferentes tipos de informação, então como escolhê-las e combiná-las? Além disso, como construir um *framework* que pode ser aplicado para qualquer cidade se informação suficiente estiver disponível? Por exemplo, é mais fácil configurar um processo de descoberta de conhecimento usando uma plataforma social (como Twitter), mas muitos aspectos da cidade e suas dinâmicas podem estar sendo perdidos.

Diversos trabalhos do estado da arte já exploraram dados geolocalizados para descobrir correlações entre a cidade e seus cidadãos. Cranshaw et al. (2010) propuseram um modelo para detectar ligações entre interações virtuais e da vida real. Hasan et al. (2013) focaram em analisar mobilidade urbana humana e padrões de atividade para compreender melhor o comportamento agregado e individual de usuários na cidade. Adicionalmente, métodos tradicionais de planejamento urbano como pesquisa de usuário e inspeções podem

[1] https://foursquare.com/

[2] https://www.instagram.com/

[3] https://www.facebook.com/

[4] https://twitter.com

ser substituídos por soluções automáticas que oferecem custo menor e mais velocidade. Nesse contexto, Frias-Martinez et al. (2014) forneceu uma técnica não-supervisionada que automaticamente determina a caracterização de tipo de uso de terreno aplicando *clustering* em regiões com padrões semelhantes de *tweets*. Similarmente, Cranshaw et al. (2012) sugeriu um modelo de *clustering* e metodologia de pesquisa para estudar a estrutura e dinâmica de cidades com base em seus residentes de maneira automática.

Apesar destes trabalhos explorarem com sucesso dados de redes sociais para descobrir informações de cidades e seus cidadãos, a sua maioria usa somente uma porção do que está disponível no ecossistema das plataformas sociais atualmente. Isso pode ser observado nas seções de trabalhos futuros de cada trabalho, onde geralmente possíveis melhorias são sugeridas ao extrair novas informações da mesma ou de outras redes sociais.

Este trabalho foca em analisar informações de diversas plataformas sociais aplicadas ao contexto de cidades inteligentes, com o objetivo de agregar diferentes conteúdos para entender melhor dinâmicas de cidades. O objetivo deste trabalho é avançar a literatura atual e testar novas formas de representação de *clusters* para avaliar como diferentes aspectos de dados sociais podem beneficiar métodos atuais, considerando que os trabalhos mencionados somente focaram em um aspecto específico dos dados. Isso é atingido experimentando diferentes dimensões de dados de redes sociais para encontrar novos pontos de vista da cidade, complementando-se para gerar uma imagem mais completa. Por exemplo, catálogos de estabelecimento como Google Places[5] e redes sociais como Facebook e Instagram possuem informações complementares que podem melhorar métodos existentes. Além disso, é assumida a hipótese que a utilização do Instagram como fonte de *check-ins* representa melhor as dinâmicas da cidade devido a sua popularidade atual e base de usuários mais diversa quando comparada com Foursquare ou redes sociais proprietárias.

A contribuição deste trabalho para a comunidade de Ciência da Computação é a extensão de técnicas de *clustering* com novas dimensões de caracterização de estabelecimentos, cada uma refletindo diferentes contextos da cidade de modo a gerar diferentes visões da estrutura e dinâmicas da cidade. Por este motivo, as contribuições se estendem também para planejadores urbanos, governos e cidadãos que possuirão mais ferramentas e informações para evoluir a cidade cada vez mais em uma cidade inteligente.

Experimentos foram executados usando uma base de dados de *check-ins* de Porto Alegre, Brasil aplicando *clustering* espectral com as representações propostas. Foi observado

[5] https://cloud.google.com/maps-platform/places/

que dependendo da característica da dimensão, o número de regiões descobertas sofreu variações e o resultado gerou uma visão diferente da cidade. Por isso, as representações propostas oferecem visões complementares da cidade quando comparados com trabalhos atuais, que contribuem com mais descobertas e fornecem uma imagem mais completa da cidade.

Esta dissertação é organizada da seguinte maneira: o Capítulo 2 explora em mais detalhes os trabalhos relacionados, o Capítulo 3 especifica o método proposto, o Capítulo 4 explica o *framework* e a ferramenta utilizados e o Capítulo 5 detalha os experimentos realizados. Finalmente, o Capítulo 6 conclui a dissertação com observações e direções futuras.

2 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados os trabalhos do estado da arte no uso de informações de redes sociais aplicado a planejamento urbano, com o objetivo de evidenciar suas contribuições e identificar pontos em aberto que sirvam de possibilidades para desdobramentos em novos trabalhos. O capítulo está organizado da seguinte forma: inicialmente são descritos os trabalhos relacionados; em seguida, os trabalhos são comparados entre si; por fim, são identificados os pontos em aberto que são contemplados com o método KANDOR proposto no capítulo 3.

2.1 Descrição dos Trabalhos Relacionados

Entre os trabalhos relacionados na área de redes sociais aplicadas a cidades inteligentes, existe uma tendência em tentar compreender o relacionamento entre localizações da cidade e atividades e preferências de seus residentes. Esta análise pode focar em pessoas que vivem na cidade, em áreas delimitadas por características específicas ou ambos. Por exemplo, é possivel analisar o comportamento de usuários, popularidade de regiões, rotinas, mapeamento e transição entre áreas e extração de pontos turísticos para compreender melhor as cidades (Silva et al., 2013). Nas seções subsequentes, são descritos os principais trabalhos relacionados ao método KANDOR apresentado no Capítulo 3. Esses trabalhos foram selecionados por abordarem com maior proximidade o método proposto.

2.1.1 Bridging the Gap between Physical Location and Online Social Network

O estudo da relação entre interações *online* e *offline* vem ganhando cada vez mais importância com a utilização constante de smartphones e aplicativos LBSN (*Location Based Social Networks*) como Foursquare e Gowalla[6]. Com isto, há cada vez mais dados a serem analisados que podem gerar novos *insights* sobre comportamento humano nas cidades. Um dos grandes desafios associados a esta área é descobrir propriedades do comportamento humano a partir dos dados geolocalizados gerados pelo usuário. Para endereçar isso, uma possível solução é analisar a co-ocorrência de localizações entre dois usuários para indicar

[6] https://en.wikipedia.org/wiki/Gowalla

possíveis associações sociais (como amizade) entre eles. Entretanto, co-localização entre pessoas desconhecidas em cidades é comum e medidas mais significativas são necessárias.

Para resolver este problema, o artigo de Cranshaw et al. (2010) propõe métricas para indicar propriedades sociais como entropia e avalia estas medidas analisando dois problemas: descobrir se dois usuários são amigos e quantos amigos um usuário tem nas redes sociais. Para realizar os experimentos, os autores utilizaram a rede social Locaccino[7] com 489 participantes da região de Pittsburgh, a fim de utilizar um ecossistema bem definido e evitar uma densidade desigual de regiões geográficas.

As principais contribuições desse artigo são: um modelo de amizade em redes sociais *online* com base em dados de co-localização; identifica-se uma correlação positiva entre padrões de mobilidade do usuário e sua quantidade de amigos; e comprova-se que medidas de diversidade da localização como a entropia dos visitantes diferentes contribuem para analisar o contexto social daquela localização.

A partir disto, utiliza-se uma rede chamada *Co-Located Friends* que representa usuários que são amigos no Facebook e possuem ocorrências de co-localização, ou seja que estiveram no mesmo local. Para diferenciar lugares mais privativos como residências de locais mais públicos como centros comerciais, são empregadas variáveis que definem a diversidade dos locais: frequência, quantidade de usuários e entropia. A frequência é o número de observações geradas no local, a quantidade de usuários é o número de usuários únicos que estiveram no local e a entropia leva em conta a distribuição das observações em relação aos usuários que estiveram ali. Isto significa que, locais com alta entropia possuem usuários com observações em proporção semelhante (centros comerciais) e locais com baixa entropia são onde as observações são concentradas em poucos usuários (como residências).

Em seguida, é analisada a correlação entre a quantidade de amigos de um usuário e as variáveis selecionadas através do cálculo da correlação de Pearson. É observado por exemplo que características pertencentes à dimensão de Diversidade do Local e Regularidade da Mobilidade possuem a maior correlação com usuários com maior quantidade de amigos. Com o intuito de entender melhor essas relações, é empregada uma análise de regressão múltipla nos dados e comprova-se que o contexto do local e mobilidade de usuários são fatores que possuem forte indicação da quantidade de amigos do usuário.

Como resultado, os autores fornecem uma ótima análise sobre os benefícios da utilização de informações sociais em *check-ins* para modelar comportamento humano e prever

[7] http://www.locaccino.org/

relacionamentos de amizades. Além disso, demonstram a importância de atributos como entropia de locais para estabelecer quantos amigos um usuário possui. Por outro lado, este trabalho possui um foco mais direcionado aos usuários ao invés dos estabelecimentos e utiliza uma rede social proprietária, o que pode gerar resultados diferentes das plataformas sociais populares.

2.1.2 An Empirical Study of Geographic User Activity Patterns in Foursquare

Com o intuito de analisar as informações presentes em redes sociais e suas possíveis aplicações para mobilidade urbana e espaços urbanos, Noulas et al. (2011) apresentam o primeiro estudo em larga-escala na LBSN mais popular da época, Foursquare. O objetivo principal dos autores é analisar aspectos espaço-temporais, distribuição das publicações por usuário e local e as possível correlações entre *check-ins* consecutivos.

Entre Maio e Setembro de 2010, foram coletados 12 milhões de *check-ins* do Foursquare (através do Twitter) gerados por 679 mil usuários. Além disso, essa base de dados foi enriquecida com informações do Facebook sobre o estabelecimento, como coordenadas geográficas, categoria, número de check-ins, visitantes únicos e endereço.

As análises feitas com os dados foram as seguintes: uma função de distribuição cumulativa complementar dos *check-ins* por local e por usuário, uma distribuição dos *check-ins* para as 10 categorias mais populares por faixas de tempo e uma análise da correlação entre *check-ins* consecutivos. Os resultados demonstraram que poucos lugares e usuários possuem uma grande quantidade de *check-ins* enquanto o restante possui uma quantidade menor. Adicionalmente, foi observado que dias da semana possuem horários definidos de pico enquanto atividades no fim de semana são mais uniformemente distribuídas durante o dia. Ao examinar as correlações, foi descoberto que para intervalos de tempo menores a probabilidade de uma visita consecutiva ocorrer para algumas atividades cresce consideravelmente.

Em conclusão, este artigo fornece uma visão geral dos diferentes tipos de informação que podem ser observados sobre mobilidade urbana usando dados do Foursquare. Os autores abordam tanto características de frequência das publicações quanto aspectos espaço-temporais para ilustrar diferentes tipos de aplicação com os dados da rede social. Apesar disso, as informações observadas fornecem conclusões que poderiam ser realizadas por um pesquisador humano, sendo somente um primeiro passo para a análise de seus potenciais.

2.1.3 The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City

Para entender as características e dinâmicas de uma cidade, normalmente são necessárias muitas pesquisas de modo que essa análise é custosa e não-escalável. Para resolver isto, Cranshaw et al. (2012) introduzem um modelo de *clustering* utilizando *check-ins* de usuários no Foursquare para descobrir padrões de atividade coletiva e assim determinar características de regiões da cidade de modo automático. O modelo proposto leva em conta tanto proximidade espacial entre locais quanto a proximidade social, derivada pela distribuição de pessoas que fazem *check-in*. A hipótese dos autores é que não somente os tipos de locais definem a região mas também as pessoas que a frequentam. Em suma, o artigo oferece três contribuições: um modelo de *clustering* que reflete os padrões de movimentos de pessoas da cidade, uma metodologia baseada em entrevistas semi-estruturadas para validação dos clusters descobertos e uma ferramenta *web* para visualização destes clusters.

Este artigo foi selecionado como *baseline* para esta dissertação pelas seguintes razões: (i) é um dos principais trabalhos do estado da arte na área de redes sociais aplicadas em planejamento urbano; (ii) oferece uma grande oportunidade para adicionar novas plataformas sociais; e (iii) foca em um problema muito interessante que é o mapeamento automático de áreas similares na cidade com base em atividades de cidadãos e suas preferências.

Para avaliar as regiões encontradas com este agrupamento, os autores conduziram entrevistas com 27 pessoas de diferentes áreas de Pittsburgh para observar quão bem seus conhecimentos da cidade se equivalem aos clusters descobertos. Possíveis aplicações vão desde algoritmos automatizados para deixar a infraestrutura de cidades mais inteligente até negócios e pesquisadores que querem entender e explorar melhor as cidades.

Os dados utilizados são uma combinação de 11 milhões de check-ins do Foursquare do *dataset* de Chen et al. (2011) e o *dataset* dos autores do artigo com 7 milhões de *check-ins* entre Junho e Dezembro de 2011. Como os check-ins do Foursquare por default não são públicos, estes foram obtidos do *timeline* público do Twitter e alinhados com a API de locais do Foursquare. Além disso, foram considerados somente *check-ins* da região de Pittsburgh.

Para descobrir as áreas urbanas, foi utilizado um algoritmo de *clustering* spectral, o qual se beneficia de uma matriz de afinidade proposta misturando afinidade espacial e social. Esta calcula, para cada local, um *cluster* com os *m* vizinhos mais próximos de acordo com similaridade de *check-ins* de usuário. Em seguida, clusters relacionados são comparados com os *check-ins* em qualquer local pertencente a um *cluster*. Como este algoritmo é o mesmo utilizado por este trabalho, ele será explicado em mais detalhes no Capítulo 3.

Com isso, foi criado um site para visualização dos resultados. Entre Novembro e Dezembro de 2011, foram conduzidas entrevistas com 27 residentes de Pittsburgh. O objetivo primário destas entrevistas era validar os *clusters* descobertos pelo algoritmo, analisando diferenças entre os observados e os bairros oficiais da cidade. Para a comparação, três padrões de dispersão foram considerados: (1) *split* - quando uma vizinhança municipal possui mais de um Livehood (região observada), (2) *spilled* - quando um cluster Livehood cruza as fronteiras de um município, e (3) *corresponding* - quando um *cluster* e as fronteiras municipais coincidem. Estes padrões refletem cada um uma dinâmica social diferente para a região descoberta.

A partir dos resultados, é possível observar uma forte correlação entre os *clusters* descobertos e a percepção das pessoas entrevistadas. Além das conhecidas fronteiras da vizinhança, fatores como características demográficas de residentes e visitantes, desenvolvimento econômico e arquitetura influenciam fortemente as divisões entre *Livehoods* e apresentam resultados mais ricos. Por exemplo, padrões *split* geralmente demonstram demografias e funções operando na mesma área, *spilled* caracteriza áreas em transição e *corresponding* sinaliza a forte influência da vizinhança sobre as interações sociais.

Em resumo, o modelo de *clustering* apresentado pelos autores contribui para analisar dinâmicas da cidade utilizando interações geradas pelas pessoas. Além de descobrir regiões já conhecidas, ele consegue observar mudanças sutis em padrões sociais e seus efeitos na cidade. Apesar disso, os resultados possuem limitações visto que os dados são agregados a partir de publicações no Twitter: o comportamento da maioria predomina e pode ocultar dados de *outliers*, as publicações em geral são atividades que se quer compartilhar e o público se limita a usuários de redes sociais geralmente de uma demografia mais jovem.

2.1.4 Understanding Urban Human Activity and Mobility Patterns using Large-scale Location-based Data from Online Social Media

Hasan et al. (2013) buscam compreender as características sociais e comportamentais de usuários de LBSNs ao analisar e agregar *check-ins* por área e por tipo de atividade. A base de dados utilizada consiste de postagens do Twitter geradas a partir de check-ins de usuários no Foursquare para três cidades dos Estados Unidos: Nova Iorque, Chicago e Los Angeles. Para cada cidade, foram extraídos seus tweets utilizando os seus limites geográficos e em média cada cidade obteve 15 mil usuários e 45 mil check-ins cada.

Uma das principais contribuições deste trabalho é a introdução da categoria de atividade como uma nova dimensão de análise, de modo a fornecer mais informações sobre a razão da escolha do local e padrões de mobilidade. Esta informação foi obtida a partir do Foursquare e contém as seguintes categorias: Casa, Trabalho, Alimentação, Entretenimento, Recreação e Compras. Para realizar as análises, células de 200x200 metros foram definidas na cidade e um *ranking* foi atribuído a elas para cada tipo de atividade dependendo do número de *check-ins* na célula. Foi comparado então a frequência de visita para a célula com seu *ranking* em cada atividade e descobriu-se que quanto pior a colocação, menor a probabilidade de um lugar ser visitado.

Em termos de popularidade por categoria de atividade, é observado que os locais mais populares variam de acordo com o propósito da visita e é verificado que os padrões de distribuição de visitas são diferentes de acordo com a atividade, onde por exemplo *check-ins* a lojas são distribuídos pela cidade enquanto os de restaurantes são mais concentrados em áreas específicas.

Com o objetivo de encontrar as características dos centros para cada atividade, foi utilizada a técnica Estimativa de Densidade *Kernel* com as distribuições de densidade probabilísticas geradas a partir dos *check-ins* agrupados por atividade e separados por intervalos de três horas. Os resultados exibiram dois grupos com um padrão diferente: um onde o centro de atividade muda de uma região para outra de acordo com o horário do dia, como alimentação e entretenimento, e outro onde o centro permanece o mesmo durante o dia inteiro, como parques e centros comerciais que possuem a capacidade de constantemente atrair pessoas.

Para encontrar os padrões individuais de mobilidade na cidade, foram analisadas duas características: temporal e frequência. A primeira se baseia em uma distribuição de visitas por horário do dia e atividade e demonstra que o horário do dia influencia no número de visitas para uma determinada atividade. Por exemplo, a atividade Alimentação possui três picos (meio-dia, seis da tarde e onze da noite) e padrões semanais exibem uma predominância de atividades de compras e recreação nos fins de semana.

Como resultado, esse trabalho demonstra que a introdução da categoria da atividade contribui para a caracterização dos padrões de mobilidade urbana e exibem importantes relações com a chance de um lugar ser visitado. Entretanto, as contribuições do artigo se limitam a analisar as distribuições e probabilidades de visita para células pré-determinadas e não oferece um meio de encontrar regiões semelhantes, por exemplo.

2.1.5 A Case Study of Active, Continuous and Predictive Social Media Analytics for Smart City

Um caso de uso possível para dados sociais é a recomendação de estabelecimentos da cidade para visitantes com base em suas preferências salvas nas redes sociais. Balduini et al. (2014) propõem um sistema de recomendação que analisa semanticamente a informação das atividades em redes sociais do visitante e sugere lugares apropriados na cidade. Seu objetivo é demonstrar durante o evento Design Week[8] de Milão a possibilidade de recomendar outros locais para visitantes vindo para a cidade para o evento.

Os autores propuseram uma arquitetura com diferentes componentes responsáveis pelas etapas de coleta e processamento de tweets e para a recomendação ao usuário. O componente chamado *Social Listener* é responsável por registrar as buscas geolocalizadas na API do Twitter e com isso receber as publicações para as áreas escolhidas.

Caso seja a primeira publicação de um usuário, suas atividades no Twitter são coletadas, extraídas em forma de entidades semânticas (com DBpedia), o perfil do usuário é construído e armazenado. Então, para cada usuário o componente *Visitor-Venue Recommender* cria predições dos dez lugares mais prováveis que o usuário possa visitar que ainda não tenha visitado e o *Visitor Engager* envia as recomendações para a pessoa pelo Twitter.

Durante a etapa de coleta dos *tweets*, algumas tarefas de pré-processamento são executadas como identificação do local do *check-in* através de quatro comparações léxicas e extração de sentimento para permitir armazenar somente os *tweets* com sentimento positivo. Para prever as futuras conexões entre visitante e local, é utilizada uma técnica de aprendizado de máquina estatístico chamada SUNS, que possui boa performance mesmo com dados esparsos e pode ser incrementalmente melhorado quando mais dados são adicionados.

As predições alcançadas pelo sistema foram comparadas com outras técnicas como Randômica, Coeficiente de Correlação de Pearson e *MostTalked*. Pearson utiliza a similaridade dos visitantes para fazer predições e *MostTalked* atribui uma ordem para recomendação a partir da frequência que os locais são mencionados em publicações. As métricas de comparação são precisão e revocação e os resultados demonstram que enquanto a técnica Randômica falha em maioria dos casos, Pearson e SUNS obtiveram performances médias e sua revocação aumenta com o tempo. A técnica *MostTalked* obteve os melhores

[8] https://fuorisalone.it/2018/

resultados e somente foi superada por SUNS em alguns casos. Por isso, se examinou também a técnica combinada de SUNS + *MostTalked* e esta teve desempenho melhor em maioria dos casos.

Apesar de normalmente a qualidade das recomendações melhorar com o aumento de quantidade de dados, observou-se que isso não ocorreu para todas atualizações dos grafos. Isto ocorre pois o desafio de prever uma visita não depende somente dos dados mas também da escassez de conexões entre usuários e locais e da estrutura da comunidade. Também é descoberto que a popularidade dos locais é o fator mais relevante para a performance enquanto similaridade entre visitantes não obteve tanto efeito.

Esse trabalho se diferencia dos demais, pois utiliza extração semântica para obter informações dos locais e não necessita que a informação da categoria esteja disponível. Além disso, o trabalho possui uma granularidade menor, pois foca somente no evento e em locais recomendados e não a nível da cidade como um todo. Por outro lado, os autores encontraram dificuldade em engajar pessoas a utilizarem a ferramenta, o que impossibilitou a avaliação das recomendações através de *feedback* de usuários.

2.1.6 Spectral Clustering for Sensing Urban Land Use Using Twitter Activity

Uma área importante no contexto de planejamento urbano é a caracterização de terrenos urbanos, responsável por detectar o tipo de terreno de acordo com seu uso. Entretanto, esta tarefa normalmente é feita através de questionários e mapeamentos com imagens de satélite, o que impõe limitações de escalabilidade, custo e confiabilidade. Por este motivo, Frias-Martinez et al. (2014) propõem uma técnica não-supervisionada baseada em dados geolocalizados (do Twitter) para detecção automática, validada com fontes externas e testada com três cidades. Para particionar o terreno em diferentes segmentos, é utilizada a técnica de *Self-Organizing Maps* (SOM), um tipo de rede neural não-supervisionada que produz uma representação bidimensional das amostras de treinamento.

Após este processo, cada segmento de terreno é caracterizado por um vetor representando a média de *tweets* postados em dias da semana e fins de semana para intervalos de 20 minutos, resultando em 144 posições. Então, *clustering* espectral é aplicado comparando semelhança entre terrenos através de semelhança de cosseno.

Três cidades foram utilizadas para avaliar este método: Londres, Manhattan e Madrid. Os objetivos da avaliação foram: analisar até que ponto o algoritmo de identificação detecta

diferentes tipos de utilização de terreno e entender o impacto da densidade de *tweets* na precisão do método.

Então, foram analisadas as classes representativas para cada cluster com suas distribuições geográficas na cidade e assim criadas hipóteses para descrever a atividade do cluster. Por exemplo, uma das regiões possui predominância de atividades em dias da semana e seus picos são em horários como 9h30, 13h e 20h30. Com isso, concluiu-se que este *cluster* representa atividades ligadas a áreas de negócio, destacando os horários de chegada no trabalho, almoço e saída do trabalho. Analogamente, apesar de algumas diferenças entre os horários de cada cidade, foi possível detectar o tipo de cada cluster descoberto.

Para validar essas hipóteses, os *clusters* foram comparados com 3 catálogos oficiais de agências da cidade tipicamente produzidos com inspeções, entrevistas e questionários. Assim, para determinar a precisão do método, foi calculada a percentagem da intersecção entre os *clusters* e os mapas oficiais de uso de terreno para cada cidade. Como resultado, é possível observar que Manhattan apresentou a melhor precisão o que indica a possível correlação com a alta densidade de *tweets* da cidade. Além disso, maioria dos *clusters* apresentaram boa precisão, onde os maiores foram comercial e negócios, variando entre 61 e 81%.

Os resultados demonstram que os *tweets* podem contribuir para modelar e compreender usos de terreno tradicionais (residencial e comercial) e descobrir novos (vida noturna) de maneira econômica e rápida. O modelo proposto pelo artigo engloba as características diferentes de dias de semana e fins de semana e de diferentes intervalos do dia, enriquecendo os resultados. É interessante observar que, além de encontrar clusters com alta porcentagem de intersecção com áreas mapeadas oficialmente, alguns destes clusters gerados não eram nem previstos anteriormente como de vida noturna. Por outro lado, uma das limitações da solução é que a identificação do tipo dos *clusters* encontrados é feita através de observações dos pesquisadores, implicando em um passo manual do método.

2.2 Comparação

Os trabalhos relacionados são comparados através de suas principais características de dados, objetivo ou implementação. As características escolhidas são: Uso de Rede Social, Quantidade de Publicações, Objetivo Principal, Método, Avaliação com Usuários e Dimensões. Os seguintes aspectos foram analisados em relação a seu tipo e presença:

- Uso de Rede Social - essa característica aponta qual (ou quais, se mais de uma) rede social é utilizada como fonte de informação. É possível observar que maioria dos trabalhos obtém informações do Twitter especialmente por ser a plataforma social mais acessível para extrair dados. Quando se observa “Foursquare/Twitter” na tabela, isto significa que foram utilizados check-ins do Foursquare mas eles foram obtidos através da API do Twitter, visto que Foursquare não fornece sua informação em larga escala devido a preocupações de privacidade. Por outro lado, Cranshaw et al. (2010) utilizou uma rede social proprietária chamada Loccacino e Hasan et al. (2013) se diferencia dos demais ao usar mais tipos de informação além de *check-ins* ao também extrair categorias de estabelecimentos do Facebook.

- Quantidade de Check-ins - um dos atributos mais variáveis foi a quantidade de check-ins nas bases de dados dos trabalhos. Apesar de não existir uma correlação direta entre a quantidade de dados e um melhor desempenho dos algoritmos, é importante possuir um número razoável de check-ins para evitar que os resultados sejam muito dependentes da amostra. Enquanto a metade dos trabalhos possui alguns milhões de publicações, Cranshaw et al. (2012), Balduini et al. (2014) e Frias-Martinez et al. (2014) possuem valores menores que 50 mil, o que pode indicar resultados menos decisivos.

- Objetivo Principal - o objetivo principal representa a categoria de problema que os trabalhos buscam oferecer soluções. A grande maioria deles endereça caracterização urbana e mobilidade de usuário, visando compreender melhor a cidade, suas regiões e como os cidadãos interagem entre si. Recomendação de estabelecimentos e predição de amizades são exemplos de contextos mais focados nos indivíduos, ajudando a mapear as interações dos usuários com seus amigos e provendo informações baseadas em suas preferências.

- Método - para atingir os objetivos propostos, são utilizados métodos estatísticos ou de aprendizado de máquina de diversas complexidades. Noulas et al. (2011) e Hasan et al. (2013) utilizam diversas técnicas de análise estatística – como função de distribuição cumulativa normal e estimativa de densidade kernel - para fazer observações sobre os dados, dependendo de uma pessoa para examinar os resultados. De maneira mais independente, Cranshaw et al. (2010) empregam classificadores, por exemplo Random

Forest e AdaBoost, e Balduini et al. (2014) aplicam Fatoração de Matrizes para seu problema de recomendação. Finalmente, Cranshaw et al. (2012) e Frias-Martinez et al. (2014) utilizam um método específico de agrupamento chamado clustering espectral para mapear as regiões da cidade.

- Avaliação com Usuários - Para avaliar problemas relacionados a redes sociais e áreas urbanas, é importante coletar as opiniões de residentes da região para obter um valor de sua aplicação real. Este atributo indica se o trabalho executou uma etapa de avaliação com usuários, geralmente com questionários online ou conduzidos por um entrevistador. Neste caso, Cranshaw et al. (2012) foram os únicos autores a realizar esta etapa enquanto os outros artigos se limitaram a realizar análises estatísticas sobre os resultados, comparando com os valores desejados.

- Dimensões - um dos fatores mais importantes para avaliar o nível de riqueza de informação empregado nas representações é classificar os tipos de dados de acordo com seu significado. Entre os artigos mencionados, todas as informações se encaixam em três grandes dimensões: co-locação, mobilidade de usuário e categoria de atividade. Co-locação possui uma conotação social e é representada pela relação criada quando dois ou mais usuários visitaram um mesmo local, contribuindo para estabelecer a diversidade ou similaridade social de um local. Mobilidade de usuário é uma categoria geral que indica padrões de check-in englobando atributos como horário e frequência de visita e transição entre locais. Finalmente, categoria de atividade informa mais sobre o propósito da visita ao local, obtida através de redes sociais como Foursquare e Facebook sobre estabelecimentos. Entre os trabalhos relacionados, Cranshaw et al. (2010) e Cranshaw et al. (2012) se destacam ao utilizar a dimensão co-locação para extrair relacionamentos sociais a partir dos check-ins. Os autores restantes utilizam informações de mobilidade de usuário como frequência e horário, porém adicionalmente Noulas et al. (2011) introduzem o uso de categoria de atividade em suas análises e Hasan et al. (2013) estende a noção dessa dimensão ao pesquisar horários de atividade das categorias de atividade. Apesar de Frias-Martinez et al. (2014) também utilizar algo parecido com essa categoria, eles não foram incluídos pois os tipos de uso de terreno são descobertos pelo algoritmo e não extraídos como fonte de dados.

2.2.1 Análise da Comparação

A Tabela 2.1 apresenta os trabalhos relacionados com as características mencionadas de modo a ilustrar melhor a comparação entre eles.

Tabela 2.1 – Comparação entre trabalhos do estado da arte

Trabalho	*Uso de Rede Social*	*Quantidade de Check-ins*	*Objetivo Principal*	*Método*	*Avaliação com Usuários*	*Dimensões*
Cranshaw et al. (2010)	Locaccino (proprietário)	2 milhões	Predição de Amizade	Classificação	N	Co-locação e Mobilidade de Usuário
Noulas et al. (2011)	Foursquare/Twitter	12 milhões	Mobilidade de Usuário/Caracterização de Regiões	Análise Estatística	N	Mobilidade de Usuário e Categoria de Atividade
Cranshaw et al. (2012)	Foursquare/Twitter	42 mil	Caracterização de Regiões	*Clustering* Espectral	S	Co-locação
Hasan et al. (2013)	Foursquare/Twitter e Facebook	1 milhão	Mobilidade de Usuário	Análise Estatística	N	Mobilidade de Usuário e Categoria de Atividade
Balduini et al. (2014)	Twitter	842	Recomendação de Estabelecimentos	Fatoração de Matrizes	N	Mobilidade de Usuário
Frias-Martinez et al. (2014)	Foursquare/Twitter	10 mil	Caracterização de Terreno Urbano	*Clustering* Espectral	N	Mobilidade de Usuário

Fonte: Rapacki (2017).

Uma das grandes limitações dos trabalhos descritos é a ausência de uma validação dos métodos com uma grande quantidade de dados e envolvendo usuários. Enquanto metade possui uma base de dados suficientemente grande para representar atividades na cidade, somente Cranshaw et al. (2012) realiza entrevistas com usuários para coletar suas opiniões sobre os resultados. Mesmo assim, foi utilizado um método qualitativo de pesquisa e por isso é dificil comparar diretamente as regiões encontradas. KANDOR endereça essa limitação construindo uma base de dados de tamanho significativo (aproximadamente um milhão de *check-ins*) e é validado através de uma pesquisa quantitativa e qualitativa com residentes da cidade.

Apesar de fornecerem soluções interessantes para os problemas, os trabalhos mencionados utilizam uma pequena variedade de tipos de informação para enriquecer seus métodos. Por exemplo, aspectos como interações sociais nas publicações (curtidas ou

comentários) ou avaliação dos estabelecimentos adicionam novas dimensões para a análise. Por esta razão, a hipótese deste trabalho é que métodos de *clustering* podem obter resultados diferentes ao incorporar novas dimensões de informação dos mesmos dados e de outras redes sociais.

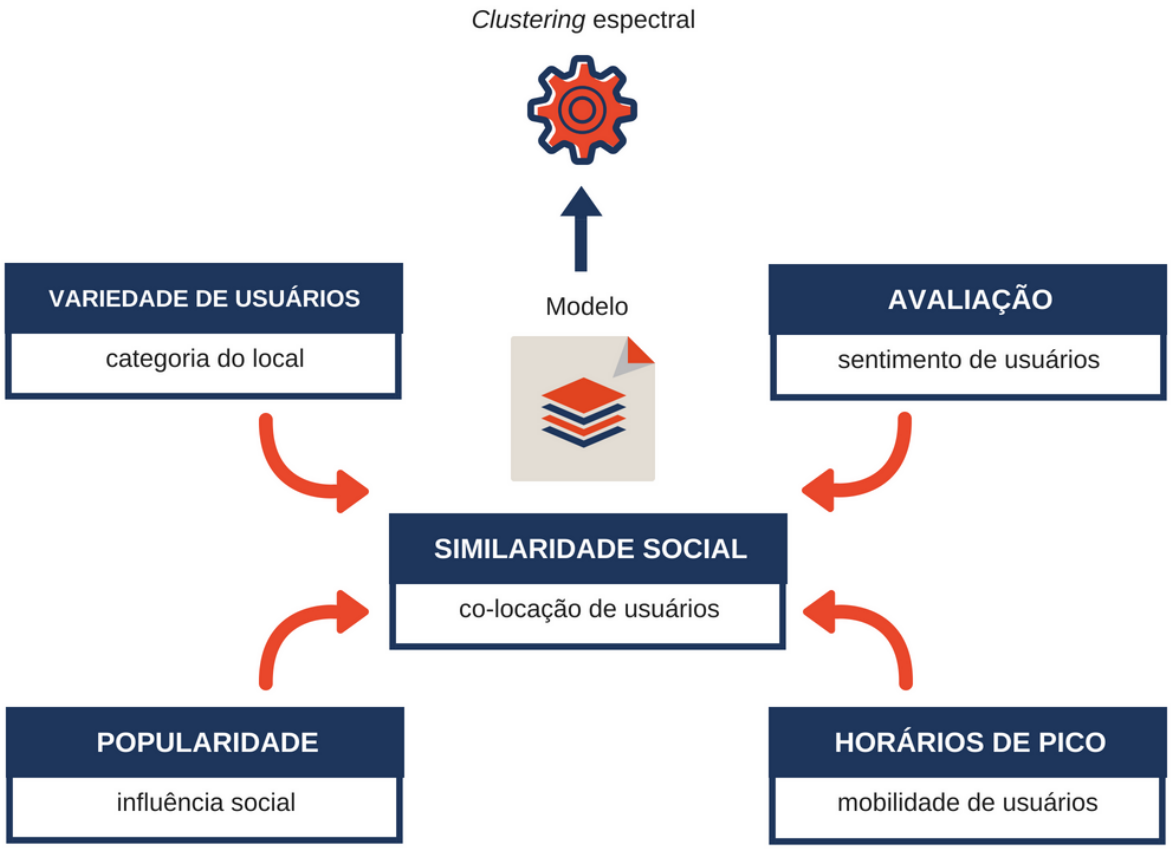
3 KANDOR – MÉTODO PARA GERAR NOVAS FORMAS DE REPRESENTAÇÃO DE CLUSTERS

Este Capítulo apresenta o método chamado KANDOR – Análise de Conhecimento de Dinâmicas em Vizinhanças e Relacionamentos *Online* (*Knowledge Analysis of Neighborhood Dynamics and Online Relationships*, em inglês). KANDOR tem como objetivo oferecer um método para gerar novas formas de representação de *clusters* a partir de dados de diferentes origens para descoberta de conhecimento em cidades inteligentes, se beneficiando da riqueza das informações presentes em redes sociais. Por este motivo, os dados são obtidos a partir de plataformas sociais para analisar como diferentes dimensões de informação, como popularidade, avaliação e horários de pico, podem contribuir para compreender melhor as características e a dinâmica das cidades.

3.1 Visão Geral

O método KANDOR tem como objetivo avançar a literatura atual e propor diferentes formas de representação de *clusters* utilizando dados de contexto e origem variados para obter visualizações mais completas das cidade. Estes contextos, que chamamos de dimensões, representam os tipos de informação que os dados representam e neste trabalho são: Similaridade Social, Variedade de Usuários, Popularidade, Avaliação e Horários de Pico. Estas cinco dimensões foram selecionadas com base nas informações disponíveis em redes sociais e por sua possível relevância para a cidade. Para cada dimensão, um modelo correspondente de *clustering* é utilizado com o algoritmo de *clustering* espectral descrito na subseção 3.3, de modo a descobrir regiões similares dentro do contexto de cada dimensão. A Figura 3.1 exibe as principais contribuições deste trabalho, destacando o algoritmo de *clustering* espectral e as dimensões utilizadas para construir representações diferentes.

Figura 3.1 – Formas de representação de *clusters* e como elas se relacionam



Fonte: Rapacki (2017).

A imagem ilustra como o algoritmo de *clustering* espectral é processado com o modelo gerado pela respectiva dimensão. Para similaridade social, as informações de *check-in* por usuário compõem o vetor de cada estabelecimento no modelo. Para as outras quatro dimensões, são adicionados atributos a este modelo base que refletem o tipo de informação da categoria em questão. Além disso, com base na tabela 2.1 dos trabalhos relacionados, uma categoria geral de dimensões foi atribuída para as dimensões escolhidas, onde duas apresentam categorias inéditas entre as referências. Para mais detalhes, as fórmulas que definem as representações para cada dimensão serão apresentadas na próxima subseção.

3.2 Dimensões

Neste trabalho, as dimensões selecionadas para análise são Similaridade Social, Variedade de Usuários, Popularidade, Avaliação e Horários de Pico. As primeiras duas foram escolhidas a partir de trabalhos relacionados e as últimas três são propostas por este trabalho para experimentação. Estas dimensões foram escolhidas das informações disponíveis nas redes sociais e com base na hipótese que cada uma pode fornecer um entendimento diferenciado do comportamento e dinâmica da cidade. Para todas as dimensões, as suas representações definem os estabelecimentos como um vetor de atributos relevantes a dimensão e todas possuem os mesmos atributos de Similaridade Social que será explicado a seguir. As próximas subseções especificam as dimensões e estão organizadas da seguinte forma: objetivo da dimensão, especificação e exemplo prático.

3.2.1 Similaridade Social

Similaridade social é uma dimensão introduzida por Cranshaw et al. (2012) e é utilizada como *baseline* por este trabalho. Ela é representada pela co-ocorrência de visitas de usuários aos mesmos locais, ou seja, se as atividades dos usuários que os frequentam são semelhantes. Desta maneira, seu objetivo é encontrar locais que costumam ser frequentados pelo mesmo grupo de pessoas e por isso mais semelhantes no contexto social.

Seu modelo, exibido na fórmula a seguir, consiste em um vetor V_i para cada estabelecimento *i*, onde o atributo $V_{i,u}$ indica o número de vezes que o usuário u fez *check-ins* no local, aqui representado por $C_{i,u}$.

$$V_i = [C_{i,1}, C_{i,2}, ..., C_{i,u}]$$

Esta dimensão compara estabelecimentos com base em quais usuários fizeram *check-ins* no mesmo local e quantas vezes o fizeram. Para esta dissertação, essa informação é obtida do Instagram, onde cada publicação representa um *check-in* de usuário associado com sua localização, se houver alguma. Por exemplo, um estabelecimento é semelhante a outro se são frequentemente visitados pelas mesmas pessoas. Em outras palavras, eles são próximos socialmente se são frequentados pelos mesmos grupos sociais.

3.2.2 Variedade de Usuários

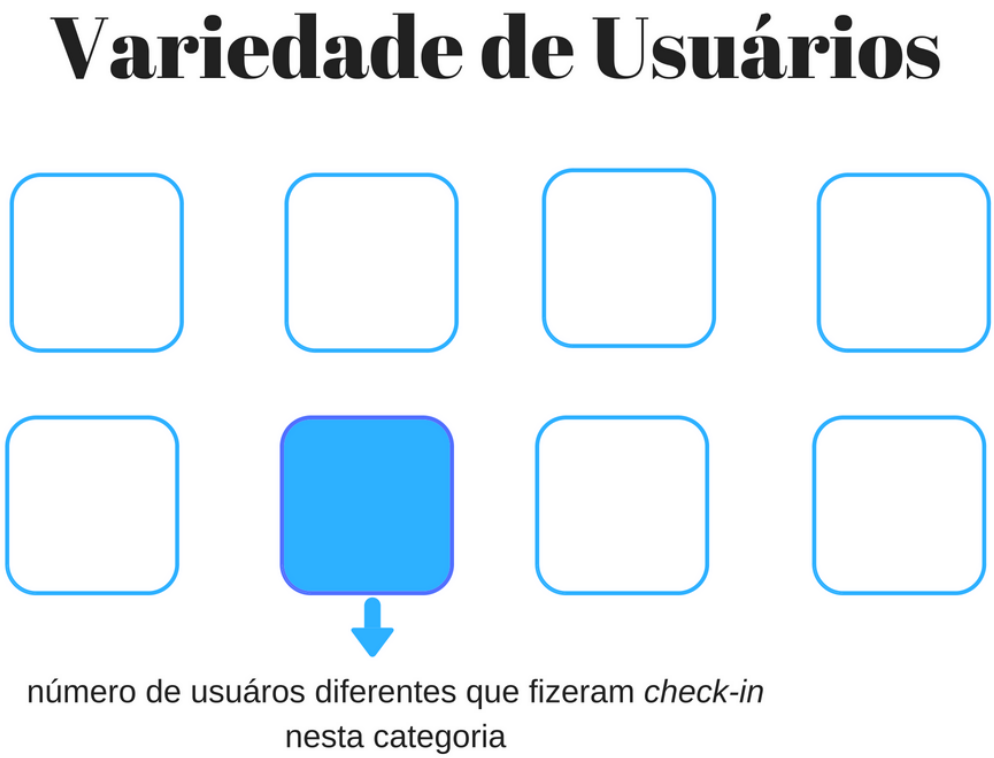
A segunda dimensão é uma variação da utilizada por Cranshaw et al. (2010) e representa a quantidade de usuários diferentes (entropia) que fizeram publicações em estabelecimentos de uma certa categoria. Deste modo, para todos os locais de uma categoria, sua entropia equivale ao número de usuários únicos que visitaram qualquer um dos locais desta categoria. Esta dimensão busca capturar o quanto o tipo de atividade da categoria consegue atrair visitantes diferentes.

O modelo desta dimensão adiciona *n* atributos ao vetor V_i mencionado acima, onde a categoria que o local pertence (E_{cat}) possui a quantidade de usuários diferentes que visitaram a categoria $_{cat}$ e o restante das *n-1* categorias possuem valor zero. A fórmula foi definida desta maneira para criar uma distância maior na etapa de calcular a semelhança de cosseno entre os vetores.

$$V_i = [C_{i,1}, C_{i,2}, ..., C_{i,u}, E_1, E_2, ..., E_{cat}, \ldots, E_n]$$

Neste contexto, nesta dissertação, as categorias são obtidas a partir da API do Facebook Places que resulta em 70 diferentes categorias na base de dados. Esta característica diferencia lugares que são visitados por muitas pessoas logo favorecendo encontros acidentais e lugares onde somente um pequeno conjunto de usuários vai. Por exemplo, um pequeno negócio possui baixa entropia enquanto um shopping center popular possui alta entropia.

Figura 3.2 – Características adicionadas ao modelo de Variedade de Usuários



Fonte: Rapacki (2017).

3.2.3 Popularidade

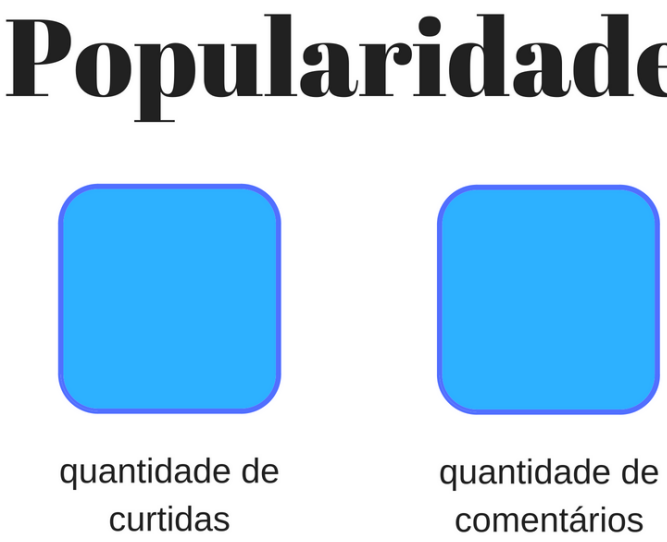
Esta é a primeira das dimensões propostas que serão avaliadas pelos possíveis benefícios que podem trazer. A popularidade de um estabelecimento é definida pela quantidade de interações sociais realizadas nas suas publicações, ou seja curtidas e comentários de seguidores. Apesar dessas interações serem publicações pessoais de usuários e poder ser relacionadas a pessoa e não ao local da publicação, pode-se considerar que lugares visitados por pessoas com alta popularidade nas redes sociais também são populares.

A fórmula para esta dimensão adiciona dois novos atributos no vetor do estabelecimento, um com a quantidade de curtidas (L_i) em todas publicações relacionadas a aquele estabelecimento e um com a quantidade de comentários (K_i) nas mesmas publicações.

$$V_i = [C_{i,1}, C_{i,2}, ..., C_{i,u}, L_i, K_i]$$

A informação sobre curtidas e comentários em publicações foi obtida do Instagram em para cada estabelecimento e atribui notas maiores para locais populares com base neste critério. Por exemplo, dois lugares como um restaurante e uma padaria com número similar de curtidas e comentários possuem maior probabilidade de estarem no mesmo cluster da cidade.

Figura 3.3 – Características adicionadas ao modelo de Popularidade



Fonte: Rapacki (2017).

3.2.4 Avaliação

A dimensão de avaliação está relacionada com a opinião de usuários sobre um lugar, portanto estabelece um sentimento geral mais positivo ou negativo dependendo do que os usuários pensam. A intenção desta análise é observar se existem regiões com estabelecimentos com uma avaliação similar e portanto se pode ser um fator determinante para caracterizar regiões.

Neste modelo, três atributos mapeiam diferentes tipos de avaliação para se ter uma análise mais completa: avaliação geral (A_i), avaliação média (a média das avaliações de cada usuário, representada por M_i) e quantidade de avaliações (Q_i). Escolheu-se usar ambos valores de avaliação geral e avaliação média pois a API usada fornece valores diferentes por um motivo desconhecido, então é interessante possuir as duas informações. Além disso, a quantidade de avaliações foi adicionada para diferenciar lugares que possuem poucas avaliações, o que pode prover uma opinião tendenciosa, e lugares avaliados por muitas pessoas e portanto uma fonte mais confiável.

$$V_i = [C_{i,1}, C_{i,2}, ..., C_{i,u}, A_i, M_i, Q_i]$$

Logo, nesta dimensão estabelecimentos com avaliações próximas possuem mais chances de estar na mesma região. Para esta dissertação, estas avaliações são fornecidas pela

API do Google Places e como explicado anteriormente, são apresentadas de duas maneiras. A primeira é a avaliação geral, que é fornecida pela API diretamente, e a avaliação média, onde foi calculada a média de todas avaliações parciais dos usuários. A principal diferença de outras dimensões é que os dados são provenientes de outra plataforma social, potencialmente apresentando uma perspectiva diferente.

Figura 3.4 – Características adicionadas ao modelo de Avaliação

Fonte: Rapacki (2017).

3.2.5 Horários de Pico

Esta dimensão identifica o tipo de atividade de regiões da cidade ao analisar em que momentos do dia os estabelecimentos são mais frequentados. Busca-se registrar os padrões das atividades nas regiões ao observar se os *check-ins* são realizados em períodos específicos do dia (de manhã, de tarde, de noite) ou então se são distribuídos em mais de um período de modo mais uniforme.

Quatro atributos são adicionados ao modelo base, cada um com o número de publicações em um local para um determinado período de tempo. Para melhor distinguir os períodos do dia, foi decidido usar quatro intervalos de seis hora cada para representar atividades de madrugada (meia-noite as 6), de manhã (6 ao meio-dia), de tarde (meio-dia as 6) e de noite (6 a meia-noite).

$$V_i = [C_{i,1}, C_{i,2}, ..., C_{i,u}, H_{i,madrugada}, H_{i,manhã}, H_{i,tarde}, H_{i,noite}]$$

O objetivo desta dimensão é representar quais períodos do dia os estabelecimentos são mais populares com base nas publicações para diferenciar lugares com diferentes tipos de atividade. Por exemplo, localizações de vida noturna possuem uma maior movimentação durante a noite enquanto restaurantes tendem a ser mais populares durante a tarde e a noite.

Figura 3.5 – Características adicionadas ao modelo de Horários de Pico

Fonte: Rapacki (2017).

3.3 Clustering Espectral

Após obter as informações da cidade e construir as representações, o algoritmo de *clustering* espectral de Cranshaw et al. (2012) é utilizado de modo a comparar o modelo do autor (similaridade social) com as outras dimensões propostas. *Clustering* espectral é um tipo de agrupamento com base em autovalores e autovetores e é utilizado por diversos trabalhos do estado da arte devido a seus benefícios. Por exemplo, ele reduz a dimensionalidade dos dados, evita ter que assumir premissas sobre a forma dos *clusters* e oferece bons resultados com um custo baixo computacional.

Ao executar essa técnica de agrupamento com o modelo de cada dimensão, o objetivo é encontrar regiões semelhantes de acordo com o critério da dimensão e comparar os resultados entre representações diferentes. Como entrada do algoritmo de *clustering*, cada estabelecimento é representado por um vetor, detalhado na subseção 3.2 para cada dimensão.

O algoritmo apresentado por Cranshaw et al. (2012) e utilizado em KANDOR é explicado na Figura 3.6 e detalhado a seguir. Considerando um conjunto *V* de n_v estabelecimentos, *U* de usuários, *C* de *check-ins* de usuários de *U* em *V* e o vetor de estabelecimento V*i*, a similaridade de cosseno entre dois estabelecimentos é definida como s(*i*,*j*) = $\frac{Vi \cdot Vj}{\|Vi\| \, [\![Vj]\!]}$, *i*,*j* ∈ *V*. Para encontrar os vizinhos mais próximos de um estabelecimento, a função N_m (v) representa os *m* estabelecimentos mais próximos a *v* de acordo com a distância geográfica *d*(*v*, ·).

Deste modo, a matriz de afinidade *A* = ($a_{i,j}$, *i*,*j* ∈ *V*) é caracterizada pela seguinte fórmula

$$a_{i,j} = \begin{cases} s(i,j) + \alpha, & \text{se } j \in N_m(V) \text{ ou } i \in N_m(V) \\ 0 & \text{caso contrário} \end{cases}$$

onde α é uma constante de valor baixo para evitar que valores degenerados não tenham nenhuma conexão.

O primeiro passo do algoritmo é calcular a matriz diagonal D onde cada elemento *i* da diagonal é a soma de todas colunas de matriz de afinidade A na linha *i*. A matriz L então é gerada com a subtração da matriz A pela D e é normalizada em uma matriz chamada L_{norm} multiplicando-a pelas matrizes de raiz quadrada inversa de D. Com os autovalores de L_{norm}, busca-se a maior diferença entre valores consecutivos para definir o valor *k* de número de *clusters* desejados.

Então, os seus *k* menores autovetores são utilizados como colunas para construir a matriz E de dimensões n_v x *k* e suas linhas são agrupadas em *k clusters* C_1,..., C_k através da técnica *k-means*. Definindo as linhas da matriz E como y_1,..., y_{nv}, obtém-se um conjunto de *clusters* em A onde se y_j pertence a C_i, então o estabelecimento *j* pertence ao cluster A_i. Além disso, os passos 8 e 9 realizam um processamento dos *clusters* gerados de modo a dividi-los em regiões conectas e evitar que eles ocupem uma porção muito grande da cidade por si só.

Figura 3.6 – Descrição do algoritmo de *clustering* espectral

Algorithm 1 *Spectral Clustering for Livehoods*
Input: V, $A = (a_{i,j})$, $G(A)$ the graph of A, k_{min}, k_{max}, τ

1: Compute diagonal degree matrix D with diagonal $(d_1, \ldots, d_{n_V})$ where $d_i = \sum_{j=1}^{n_V} a_{i,j}$.
2: $L := D - A$
3: $L_{norm} := D^{-1/2} L D^{-1/2}$
4: Let $\lambda_1 \leq \ldots \leq k_{max}$ be the k_{max} smallest eigenvalues of L_{norm}. Set $k = \arg\max_{i=k_{min},\ldots,k_{max}-1} \Delta_i$ where $\Delta_i = \lambda_{i+1} - \lambda_i$.
5: Find the k smallest eigenvectors $e_1, \ldots, e_k$ of L_{norm}.
6: Let E be an $n_V \times k$ matrix with e_i as columns.
7: Let the $y_1, \ldots, y_{n_V}$ be the rows of E, and cluster them into $C_1, \ldots, C_k$ with k-means. This induces a clustering on $A_1, \ldots, A_k$ by $A_i = \{j | y_j \in C_i\}$.
8: For each A_i, let $G(A_i)$ be the subgraph of $G(A)$ induced by vertices A_i. Split $G(A_i)$ into connected components. Add each component as a new cluster, removing $G(A_i)$.
9: Let b the area of bounding box containing coordinates in V, and b_i be the area of the box containing A_i. If $b_i/b > \tau$, delete cluster A_i, and redistribute each $v \in A_i$ to the closest A_j under single linkage distance $d(v, A_j)$.

Fonte: Cranshaw et al. (2012).

Os parâmetros escolhidos por Cranshaw et al. (2012) e também neste algoritmo são $m = 10$, $\alpha = 0.01$, $k_{min} = 30$, $k_{max} = 45$, e $\tau = 0.4$.

3.4 Tabela Comparativa com Trabalhos Relacionados

A Tabela 3.1 exibe uma comparação entre os principais atributos dos trabalhos relacionados e de KANDOR para enfatizar os pontos fortes do método proposto. Pode-se observar que KANDOR utiliza uma variedade maior de redes sociais e dimensões quando comparado a outros trabalhos, o que correlaciona com o objetivo da dissertação. Além disso,

KANDOR se enquadra na categoria de trabalhos que utilizam métodos de aprendizado de máquina, no caso de *clustering* espectral. A base de dados utilizada possui um tamanho considerável em relação a média dos trabalhos relacionados e KANDOR se destaca por realizar uma avaliação com usuários, o que a maioria dos outros trabalhos não fez.

Tabela 3.1 – Comparação entre trabalhos do estado da arte

Trabalho	*Uso de Rede Social*	*Quantidade de Check-ins*	*Objetivo Principal*	*Método*	*Avaliação com Usuários*	*Dimensões*
Cranshaw et al. (2010)	Locaccino (proprietário)	2 milhões	Predição de Amizade	Classificação	N	Co-locação e Mobilidade de Usuário
Noulas et al. (2011)	Foursquare/Twitter	12 milhões	Mobilidade de Usuário/Caracterização de Regiões	Análise Estatística	N	Mobilidade de Usuário e Categoria de Atividade
Cranshaw et al. (2012)	Foursquare/Twitter	42 mil	Caracterização de Regiões	*Clustering* Espectral	S	Co-locação
Hasan et al. (2013)	Foursquare/Twitter e Facebook	1 milhão	Mobilidade de Usuário	Análise Estatística	N	Mobilidade de Usuário e Categoria de Atividade
Balduini et al. (2014)	Twitter	842	Recomendação de Estabelecimentos	Fatoração de Matrizes	N	Mobilidade de Usuário
Frias-Martinez et al. (2014)	Foursquare/Twitter	10 mil	Caracterização de Terreno Urbano	*Clustering* Espectral	N	Mobilidade de Usuário
KANDOR	Foursquare, Instagram, Facebook e Google Places	850 mil	Caracterização de Regiões	*Clustering* Espectral	S	Co-locação, Mobilidade de Usuário, Categoria de Atividade, Influência Social e Sentimento de Usuários

Fonte: Rapacki (2017).

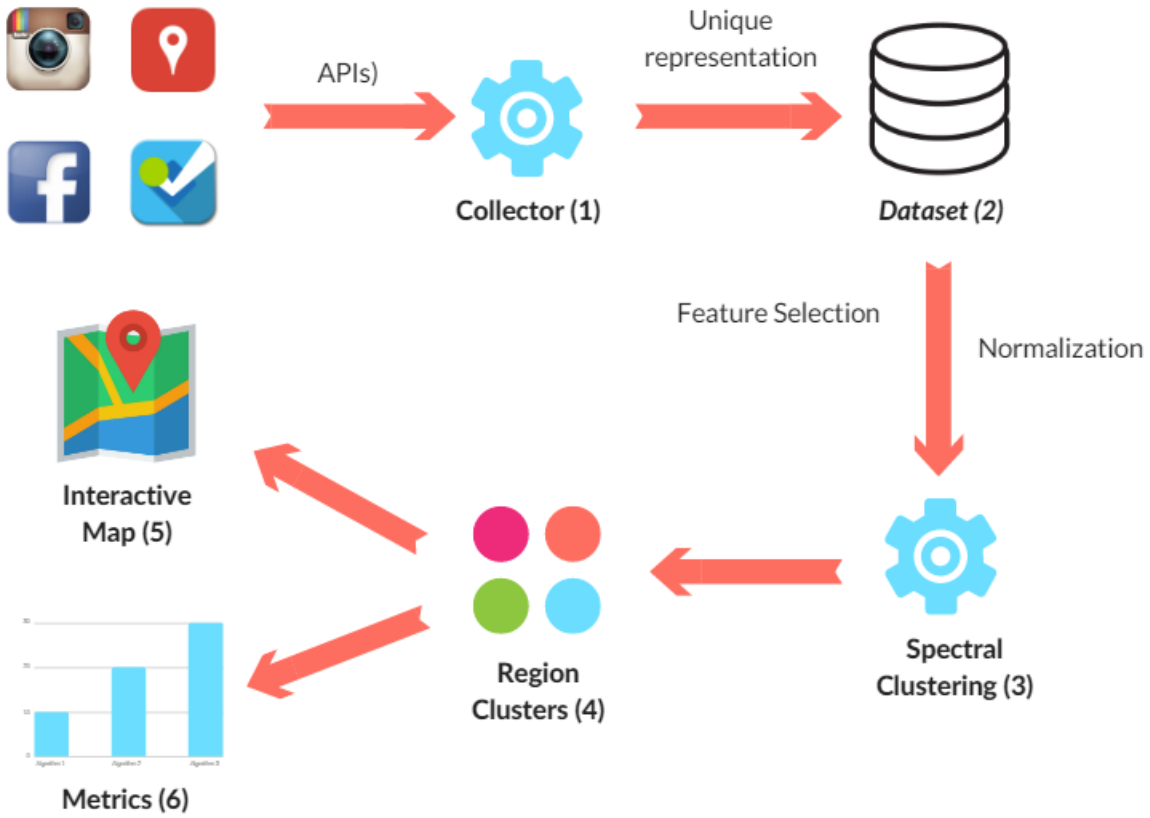
4 KANDOR: FRAMEWORK E FERRAMENTA

De modo a executar o método proposto na seção anterior, foi necessário construir uma base de dados com informações de diferentes plataformas sociais e ferramentas para ajudar a avaliação e comparação das representações propostas. Por isso, foi criado um *framework* com componentes responsáveis por diferentes etapas do processo, como coleta, processamento e visualização. As próximas seções estão organizadas da seguinte maneira: o *framework* e seus componentes são descritos em detalhes na seção 4.1 e a ferramenta *web* para visualização é apresentada na seção 4.2.

4.1 *Framework*

O objetivo do *framework* é auxiliar a execução do método KANDOR, estabelecendo componentes responsáveis por coletar e preparar os dados, executar o algoritmo de *clustering* espectral e gerar visualizações e métricas a partir dos resultados. A Figura 4.1 ilustra a visão geral dos componentes e as operações que compõem este framework. Primeiramente, um passo de extração constrói a base de dados que será usada pelo algoritmo de *clustering* espectral, onde o Coletor (1) realiza chamadas às APIs das redes sociais e cria uma representação única para cada estabelecimento.

Figura 4.1 – Visão geral dos principais componentes e operações do *framework*



Fonte: Rapacki (2017).

Em seguida, KANDOR está pronto para executar o algoritmo para descobrir regiões similares com base na base de dados (2) obtida. Antes de cada execução, uma etapa de pré-processamento obtém a informação da base de dados, aplica seleção de características com base nas configurações do algoritmo e normaliza as caracteristícas. Isto gera vetores representando estabelecimentos e uma matriz de afinidade com similaridade entre estabelecimentos.

O componente do *clustering* espectral (3) executa o algoritmo descrito no subcapítulo 4.1.4 e retorna um conjunto de *clusters* (4) que representam regiões com estabelecimentos mais similares entre si, de acordo com a configuração de similaridade do algoritmo. Estas informações são exibidas em um mapa interativo (5), possibilitando ao usuário visualizar as regiões da cidade e selecionar qual dimensão utilizar para definir a similaridade entre os estabelecimentos. Além disso, estas regiões podem ser analisadas por algoritmos e gerar métricas (6) para avaliar a resposta de cada modelo. A seguir, cada um dos componentes é descrito em mais detalhes.

4.1.1 Extração de Dados e Coletor

Para gerar uma representação completa dos estabelecimentos da cidade, diversas redes sociais são utilizadas como fonte de dados: Instagram, Google Places, Facebook e Foursquare. Através de suas APIs, o Coletor executa um *script* em Python para coletar todas as informações disponíveis sobre a cidade. Mais especificamente, os bairros são obtidos do Foursquare, informações sobre locais são obtidos do Facebook (categoria) e Google Places (avaliação) e *check-ins* são coletados a partir de publicações do Instagram, junto com seus comentários e curtidas.

Para obter amostrar heterogêneas dos estabelecimentos da cidade, uma lista do Foursquare com os 61 bairros mais centrais (de 81 no total) de Porto Alegre foi usada e o centro de cada bairro foi o ponto de partida para as buscas. Em seguida, a API do Facebook é utilizada para coletar todos locais do Facebook em um raio de 2km do centro de cada bairro, visto que o tamanho médio de um bairro em Porto Alegre possui esse tamanho. A partir desses locais, busca-se extrair representações únicas dos estabelecimentos ao combinar as informações presentes em outras redes sociais.

Utilizando o identificador de cada local do Facebook com a API do Instagram, é possível recuperar a representação correspondente no Instagram e todas as publicações compartilhadas de modo público neste local. Essas publicações contém informações como usuário que compartilhou, texto da publicação e interações sociais como curtidas e comentários na foto. Finalmente, com a localização do estabelecimento do Instagram, a API do Google Places é consultada e o lugar geograficamente mais próximo ao do Instagram é escolhido para ser a representação do Google Places. Com isso, dados como avaliação de usuários e endereço também são coletados.

Para adicionar mais redes sociais como fonte de dados, basta alterar o Coletor para executar as APIs da rede e fazer as conexões necessárias para manter uma entidade única para cada estabelecimento. As informações mencionadas são salvas e formam a base de dados que será usada nos experimentos.

4.1.2 Base de Dados

Por simplicidade, foi escolhido o banco de dados orientado a documentos MongoDB para armazenar os dados, visto que para *check-ins* e estabelecimentos não é necessário a complexidade de bancos de dados relacionais. A base consiste principalmente de informações

de publicações, detalhes dos estabelecimentos para cada rede social e os resultados dos experimentos, como os *clusters* gerados e as métricas de avaliação. A Tabela 4.1 exibe as coleções que compõem a base de dados e uma breve descrição sobre elas e a Tabela 4.2 apresenta a quantidade de dados no geral.

Tabela 4.1 – Coleções da base de dados e seus conteúdos

Coleção	*Descrição*
clusters	regiões descobertas pelo algoritmo
google_places	informações de avaliação e endereço do estabelecimento (Google Places)
insta_places	nome, geolocalização e *ids* do Facebook e Instagram
media	publicações do Instagram com usuário, texto, comentários, curtidas, etc.
metrics	avaliações de coesão-separação por *cluster*
neighborhood_polygons	áreas dos bairros da cidade
neighborhoods	bairros da cidade
places	informações de localização e categoria do estabelecimento (Facebook)
venue_checkins	vetores de estabelecimentos com *check-ins* por usuário

Fonte: Rapacki (2017).

Tabela 4.2 – Base de Dados de KANDOR

Parâmetros	*Quantidade*
Estabelecimentos	2.587
Publicações	850.695
Usuários	148.051

Fonte: Rapacki (2017).

Para representar os *check-ins*, assume-se que as publicações do Instagram com com informação de localização são mais relevantes que *check-ins* de outras redes sociais como Foursquare por três grandes motivos: Instagram possui uma “assinatura cultural” mais distinguível, é menos suscetível a mudanças ao longo do tempo (Silva et al., 2013) e nos últimos anos se tornou mais popular com um crescimento maior de usuários.

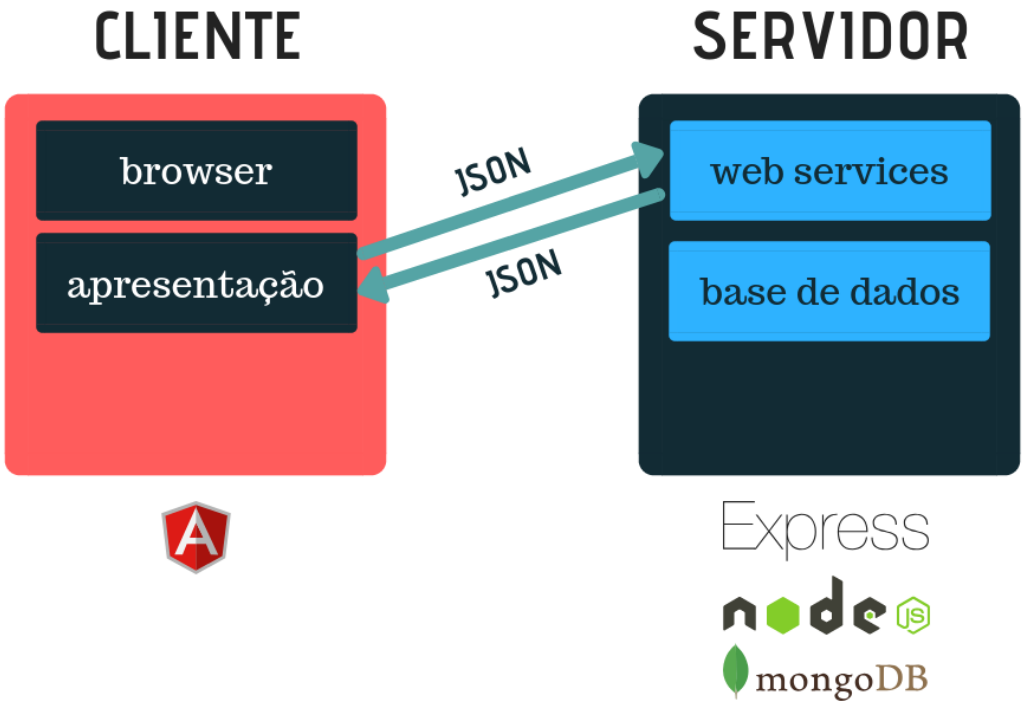
4.1.3 *Clustering* Espectral

O algoritmo de *clustering* espectral é explicado em detalhes na seção 3.2. Esse código foi implementado na linguagem Python e diversas bibliotecas foram utilizadas para contribuir com os cálculos e para salvar resultados intermediários, entre elas numpy, scipy e h5py.

4.2 KANDOR: Ferramenta de Visualização

De modo a avaliar a relevância das dimensões apresentadas e descobrir quais são as mais úteis para gerar visões da cidade, foi criada uma ferramenta *web* com a plataforma MEAN (MongoDB, Express, AngularJS e NodeJS) e a API do Google Maps para apresentar os resultados em um mapa interativo. Essa arquitetura é exibida na Figura 4.2.

Figura 4.2 – Arquitetura da página *web* para visualização



Fonte: Rapacki (2017).

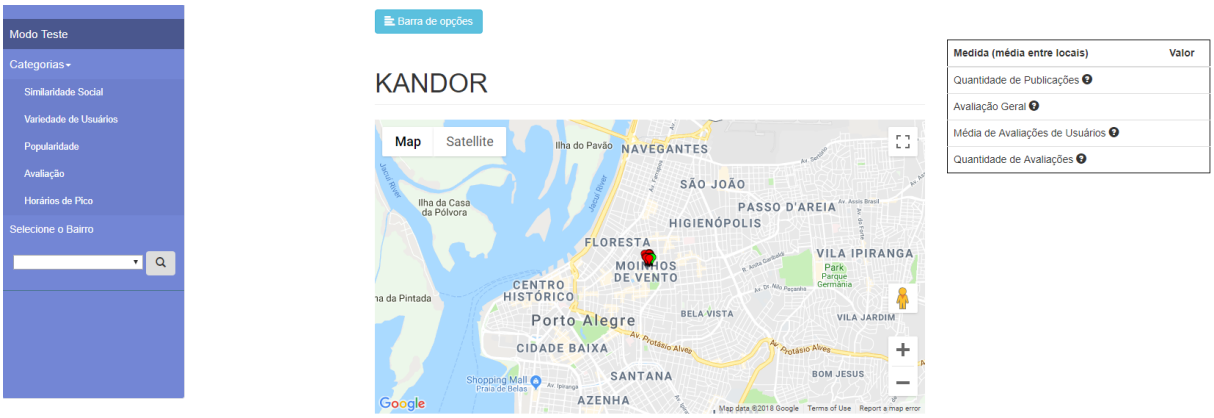
Quando uma dimensão é escolhida no KANDOR, o mapa interativamente exibe as regiões da cidade de acordo com a dimensão selecionada. Ao clicar em um local específico, a região a que ele pertence e os outros locais semelhantes são destacados e uma tabela mostra características daquela região. É possível também selecionar na barra da esquerda um bairro da cidade para destacá-lo no mapa, facilitando a comparação entre as regiões encontradas pelo KANDOR e os bairros oficiais.

A Figura 4.3 exibe uma visão geral da cidade, onde regiões com locais similares são representados com a mesma cor. A esquerda do mapa, uma barra auxilia a navegação do

usuário, permitindo trocar a dimensão (similaridade social, variedade de usuários, etc.) ou selecionar um bairro para destacar. A sua direita, está uma tabela com medidas relacionadas a características dos locais da região selecionada, como quantidade de usuários, de curtidas e avaliações. Os valores apresentados nessa tabela são relativos a média geral, ou seja indicam se o valor da região é baixo, médio ou alto em relação a todas regiões encontradas na cidade. Ao clicar em um ponto no mapa, sua região correspondente é destacada como pode ser observado na Figura 4.4, incluindo todos locais similares da mesma região.

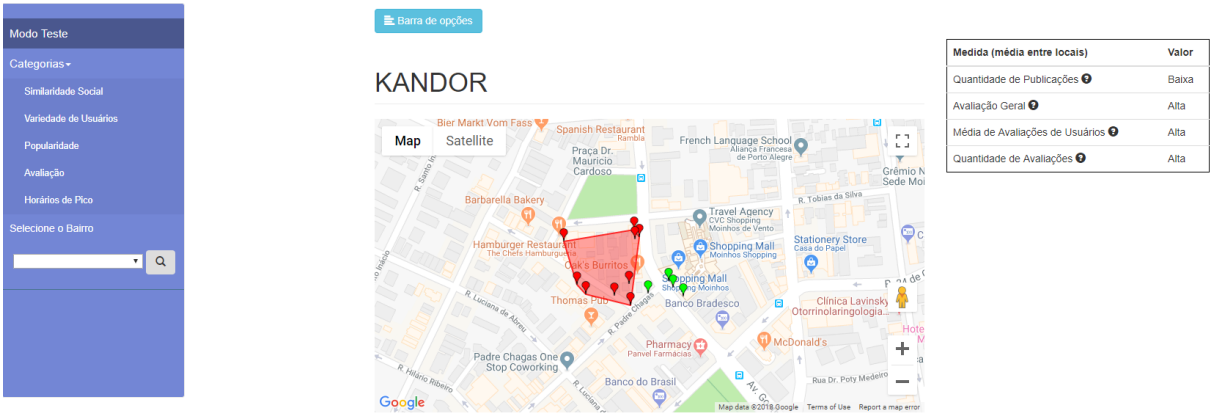
Esta visualização serve de guia para a realização dos experimentos do próximo capítulo, onde usuários respondem um questionário enquanto navegam no site para ver as regiões com atividade semelhante para cada dimensão. Desta maneira, os usuários podem comparar resultados e avaliar a informação descoberta por KANDOR.

Figura 4.3 – Página principal exibindo uma visão geral da cidade



Fonte: Rapacki (2017).

Figura 4.4 – *Cluster* destacado após usuário clicar em qualquer estabelecimento que pertence a ele

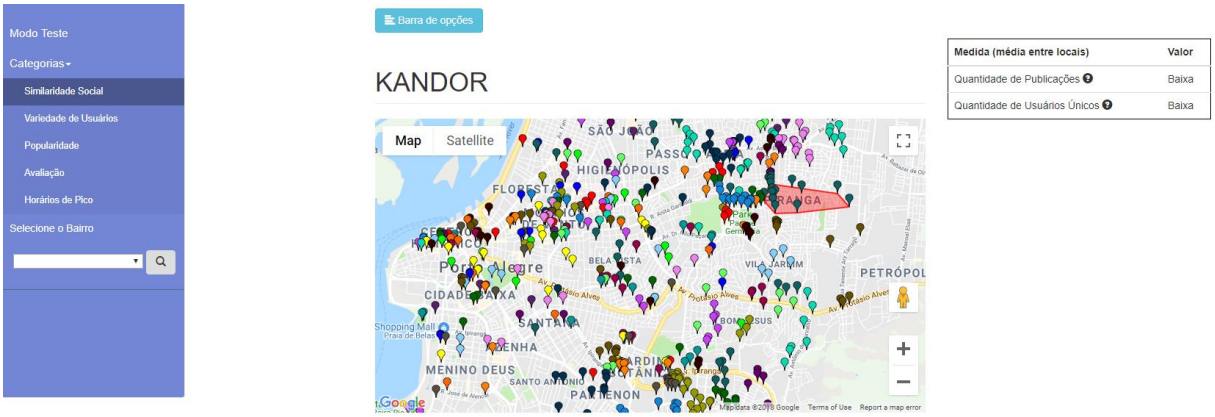


Fonte: Rapacki (2017).

4.2.1 Interação 1: Similaridade Social

A dimensão Similaridade Social é definida por lugares frequentemente visitados pelas mesmas pessoas, ou seja que há uma semelhança entre os grupos de pessoas que os visitam e em sua frequência. A Figura 4.5 exibe a visualização da página quando a dimensão Similaridade Social é selecionada. É possível observar no topo direito um polígono vermelho indicando uma das regiões selecionadas e que a tabela na direita indica que quantidade de publicações e usuários únicos para essa região é baixa.

Figura 4.5 – Mapa quando dimensão Similaridade Social está selecionada



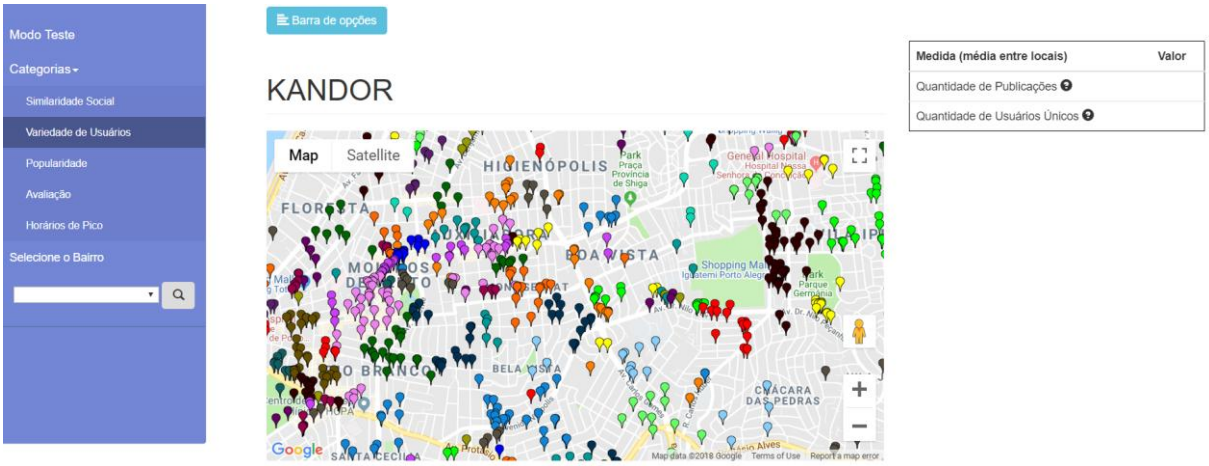
Fonte: Rapacki (2017).

A tabela à direita exibe as medidas relevantes para a dimensão, ou seja quantidade de publicações e quantidade de usuários únicos. Estes valores são importantes para o entendimento do usuário mas influenciam indiretamente no agrupamento, dado que o valor real está quando se encontram os mesmos subgrupos de usuários que frequentemente visitam o local.

4.2.2 Interação 2: Variedade de Usuários

Variedade de Usuários é representada pela quantidade de usuários diferentes que já visitaram pelo menos uma vez um local. Em outras palavras, ela mede a capacidade do local de atrair pessoas diferentes. Esta dimensão é ilustrada na Figura 4.6 quando Variedade de Usuários está selecionada. Nesta imagem, nenhuma região está selecionada logo uma visão mais geral da cidade é exibida.

Figura 4.6 – Mapa quando dimensão Variedade de Usuários está selecionada



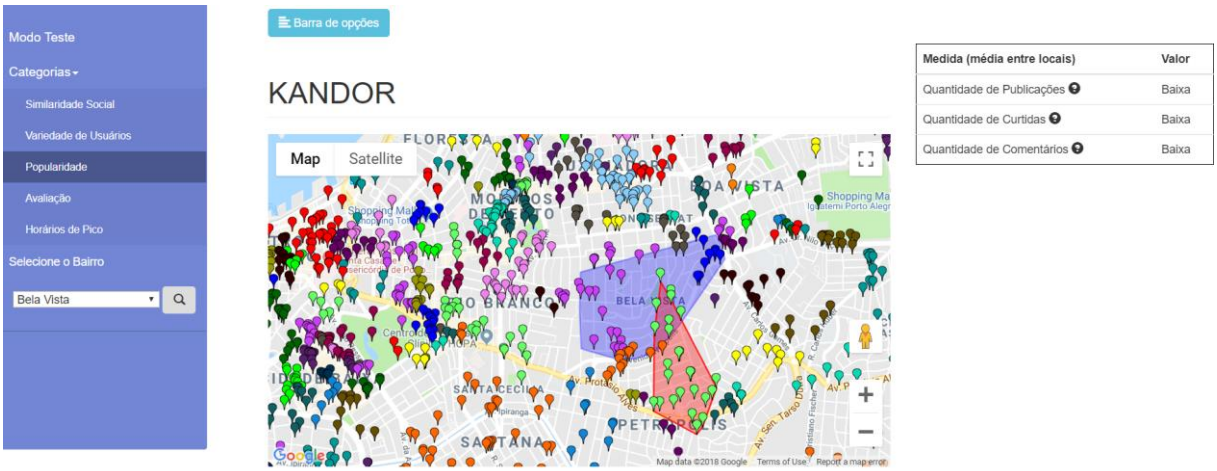
Fonte: Rapacki (2017).

Na barra de medidas, são mostradas as mesmas de Similaridade Social, mesmo que as dimensões sejam diferentes. Isto ocorre porque esses atributos são importantes para ambas, mas de maneiras diferentes. Para Variedade de Usuários, o atributo de quantidade de publicações é adicionado à quantidade de usuários únicos, que dá mais relevância para lugares com maior variedade de usuários.

4.2.3 Interação 3: Popularidade

Na Figura 4.7, pode-se observar a visualização gerada para a dimensão Popularidade, que caracteriza locais com o mesmo padrão de atividade social. Essa informação é extraída a partir das publicações de usuários nos lugares em questão, considerando mais populares as localizações com maior número de curtidas e comentários nessas publicações. A imagem mostra um destaque no bairro Bela Vista (em azul) e em uma região com locais no bairro (em vermelho).

Figura 4.7 – Mapa quando dimensão Popularidade está selecionada



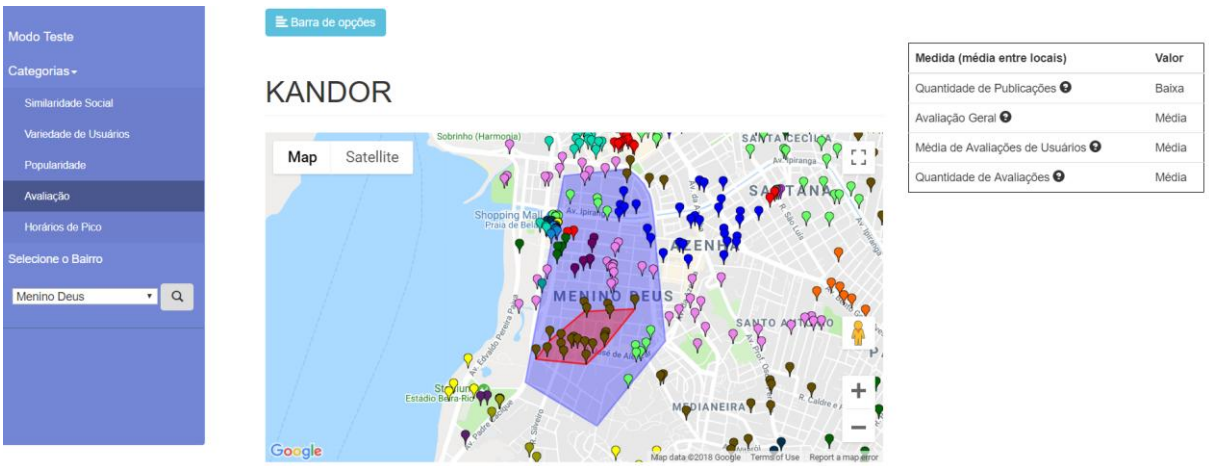
Fonte: Rapacki (2017).

Pode ser observado no menu à direita que além de quantidade publicações, são adicionadas também as medidas de quantidade de curtidas e quantidade comentários, refletindo a descrição da dimensão.

4.2.4 Interação 4: Avaliação

A dimensão Avaliação representa o sentimento dos usuários em relação aos locais, coletado a partir de avaliações da API Google Places. Deste modo, regiões são formadas conforme a semelhança entre as notas e número de avaliações que receberam.

Figura 4.8 – Mapa quando dimensão Avaliação está selecionada



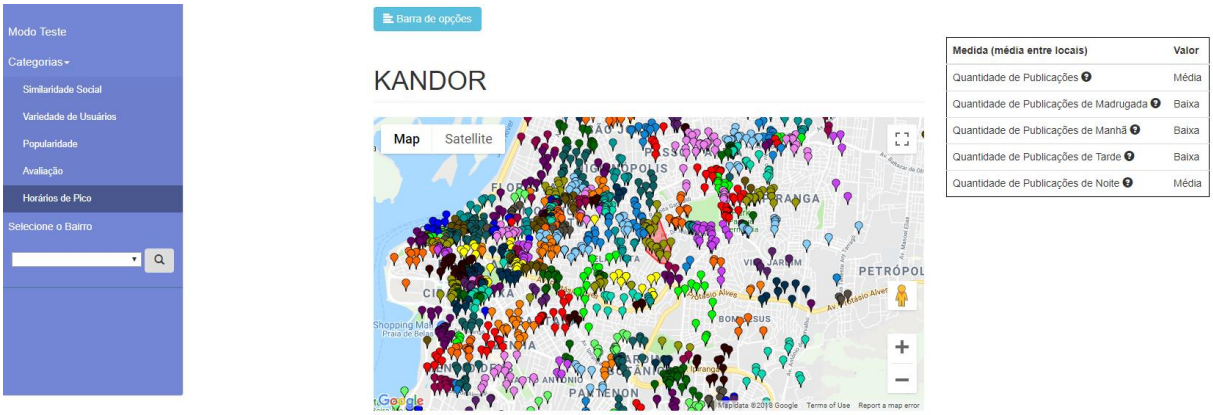
Fonte: Rapacki (2017).

A Figura 4.8 ilustra uma possível visualização da dimensão Avaliação, mostrando o bairro Menino Deus (em azul) e uma região dentro do bairro (em vermelho) como selecionados. A tabela na direita exibe as métricas relacionadas a avaliação geral do estabelecimento, a média entre todas avaliações de usuários e a quantidade de avaliações. Esse último atributo pode contribuir para indicar locais que receberam poucas avaliações e por isso sua nota não é tão indicativa de sua qualidade.

4.2.5 Interação 5: Horários de Pico

A dimensão Horários de Pico, exibida na Figura 4.9, caracteriza os locais de acordo com a distribuição de visitações durante o dia no lugar, buscando diferenciar o tipo de atividade e motivo de visita. Por exemplo, locais podem ser mais visitados durante um turno específico, como universidades durante a manhã e tarde, ou possuir uma distribuição de atividade mais uniforme como áreas comerciais. É possível observar na figura uma visão ampla da cidade através da dimensão Horários de Pico e com uma região selecionada no centro em vermelho.

Figura 4.9 – Mapa quando dimensão Horários de Pico está selecionada



Fonte: Rapacki (2017).

As medidas exibidas na tabela à direita são a quantidade relativa de publicações durante um turno de seis horas: madrugada, manhã, tarde ou noite.

5 AVALIAÇÃO

Este capítulo descreve dois conjuntos de experimentos que avaliam o método proposto. O primeiro exerimento tem por objetivo avaliar as dimensões propostas no KANDOR comparando as regiões descobertas pelo seu respectivo modelo de *clustering* e suas características. A suposição é que cada dimensão pode oferecer visões diferentes, porém complementares, com base no tipo de informação que se está analisando. Deste modo, para realizar uma avaliação mais completa das contribuições de KANDOR, foram utilizados dois meios: um método estatístico para uma primeira observação dos resultados e entrevistas quantitativas para coletar as opiniões de residentes da cidade.

Como método estatístico, é utilizada uma técnica de avaliação de *clusters* para examinar os resultados e encontrar possíveis indicações sobre as peculiaridades de cada dimensão. O segundo experimento avalia com usuários o potencial de cada dimensão para encontrar regiões relevantes da cidade através de suas percepções e conhecimentos da região. De modo a medir as possíveis contribuições no mundo real, o experimento apresentado na seção 5.2 realiza questionários com residentes da cidade através de um mapa interativo exibindo as regiões encontradas por dimensão para ajudar a visualização das diferenças, semelhanças e aspectos relevantes.

5.1 Avaliação das dimensões propostas no KANDOR

Para calibrar os parâmetros do algoritmo de *clustering* e avaliar os *clusters* gerados por cada dimensão, a quantidade de regiões e seus coeficientes de coesão-separação (Aliguliyev, 2009) foram calculados e analisados. Essa técnica de avaliação foi escolhida pois permite examinar a dispersão *intra-clusters* (entre estabelecimentos da mesma região) e a separações dos *clusters*. As próximas subseções estão organizadas da seguinte forma: a subseção 5.1.1 especifica o algoritmo da validação interna de *cluster*, a subseção 5.1.2 explica a calibragem feita no algoritmo de *clustering* e a subseção 5.1.3 apresenta os resultados da análise das dimensões.

5.1.1 Coeficiente Coesão-Separação

O coeficiente de coesão-separação de Aliguliyev (2009) combina os atributos de coesão e separação para gerar uma avaliação única, onde um valor maior representa clusters

com estabelecimentos similares entre si e distintos de outros clusters. Coesão indica a similaridade entre estabelecimentos do mesmo cluster e separação representa o grau de difereça de um cluster para outros. Dado *U* o conjunto de usuários, *C* o conjunto de clusters gerados, *V* o conjunto de n_v vetores de estabelecimento e o vetor de estabelecimento V_i, a similaridade de cosseno entre dois estabelecimentos é definida como $s(i,j) = \frac{Vi \cdot Vj}{\|Vi\| \, [\![Vj]\!]}$, $i,j \in V$. Além disso, é criada uma representação vetorial de clusters C_i similar a V_i, contendo a soma dos atributos de mesmo índice de todos estabelecimentos em sua região. C_i pode ser representado por

$$C_{i,j} = \sum_{k=0}^{n_v} V_{k,j},\ i \in C,\ j \in U,\ k \in V.$$

A fórmula do coeficiente é apresentada na Figura 5.1, onde observa-se que o algoritmo basicamente divide a média das similaridades mínimas entre estabelecimentos no mesmo cluster pela maior similaridade entre clusters. Para calcular o numerador para um cluster C_i de tamanho n_c, para cada estabelecimento V_j pertencente a C_i, encontra-se a menor similaridade entre V_j e qualquer outro estabelecimento em C_i e a média entre todas similaridades mínimas é medida. Para o denominador, busca-se a maior similaridade entre qualquer cluster em C_i e outro diferente dele.

Figura 5.1 – Fórmula para calcular o coeficiente de coesão-separação de uma coleção de *clusters*

$$CS_1(k) = \frac{\sum_{p=1}^{k} \left\{ \frac{1}{|C_p|} \sum_{D_i \in C_p} \min_{D_l \in C_p} \{sim(D_i, D_l)\} \right\}}{\sum_{p=1}^{k} \left\{ \max_{\substack{q=1,\ldots,k \\ q \neq p}} \{sim(O_p, O_q)\} \right\}}.$$

Fonte: Aliguliyev (2009).

5.1.2 Calibragem do Algoritmo de *Clustering*

Quando o método *baseline* foi executado com a base de dados de Porto Alegre, foi observado que as regiões encontradas consistiam de poucos estabelecimentos em geral. Como o algoritmo original de Cranshaw et al. (2012) é aplicado em Pittsburgh, decidiu-se realizar um experimento com um dos parâmetros de entrada de modo a calibrar os resultados.

No algoritmo de *clustering* espectral, o parâmetro *m* representa o número de vizinhos mais próximos geograficamente que se deseja associar para cada local na matriz de afinidade. Como este parâmetro possui o potencial de ser o o que mais influencia o algoritmo visto que

ele pode restringir ou expandir a "área de busca", decidiu-se experimentar diferentes valores de *m* para analisar o valor ideal para Porto Alegre.

Foram escolhidos os valores 10, 13 e 15 com o intuito de aumentar a quantidade de vizinhos mais próximos e assim gerar regiões com mais estabelecimentos. O algoritmo de *clustering* espectral foi executado com esses valores e os *clusters* descobertos foram avaliados considerando o coeficiente de coesão-separação que combina coesão interna do *cluster* e separação entre os *clusters*.

A Tabela 5.1 ilustra a grande diferença encontrada para o coeficiente de coesão-separação entre os parâmetros selecionados, onde *m*=10 alcança um valor significantemente melhor. Além disso, pode-se notar que quanto maior o valor de *m*, menos *clusters* são encontrados. Isso está de acordo com a idéia de que valores maiores comparam mais estabelecimentos vizinhos na matriz de afinidade e por isso é capaz de conectar regiões maiores. Como o valor *m*=10 obteve um coeficiente muito superior, o parâmetro foi mantido igual ao original e portanto todos os parâmetros utilizados nestes experimentos são idênticos aos do experimento de Cranshaw et al. (2012).

Tabela 5.1 – Comparação entre os valores do parâmetro *m*, número de clusters gerados e o coeficiente coesão-separação resultante

M	*Número de clusters*	*Coeficiente coesão-separação*
10	287	41.25
13	249	2.45
15	193	3.36

Fonte: Rapacki (2017).

5.1.3 Resultados da Avaliação das Dimensões

De modo a realizar uma avaliação preliminar das dimensões, foi decidido analisar as regiões descobertas utilizando o mesmo coeficiente de coesão-separação que foi usado para escolher o parâmetro *m*. Neste caso, ele pode ser usado para comparar os resultados e tentar compreender as diferentes características entre as representações com base na avaliação dos *clusters*.

Além disso, o número de *clusters* descobertos foi utilizado como comparação considerando seu impacto na visualização e entendimento da cidade. Por exemplo, exibir muitas regiões menores pode dificultar a visualização das informações e falhar em fornecer uma visão geral útil enquanto exibir poucas regiões grandes pode perder importantes detalhes de regiões menores. Enquanto em geral busca-se melhor coeficientes para gerar regiões

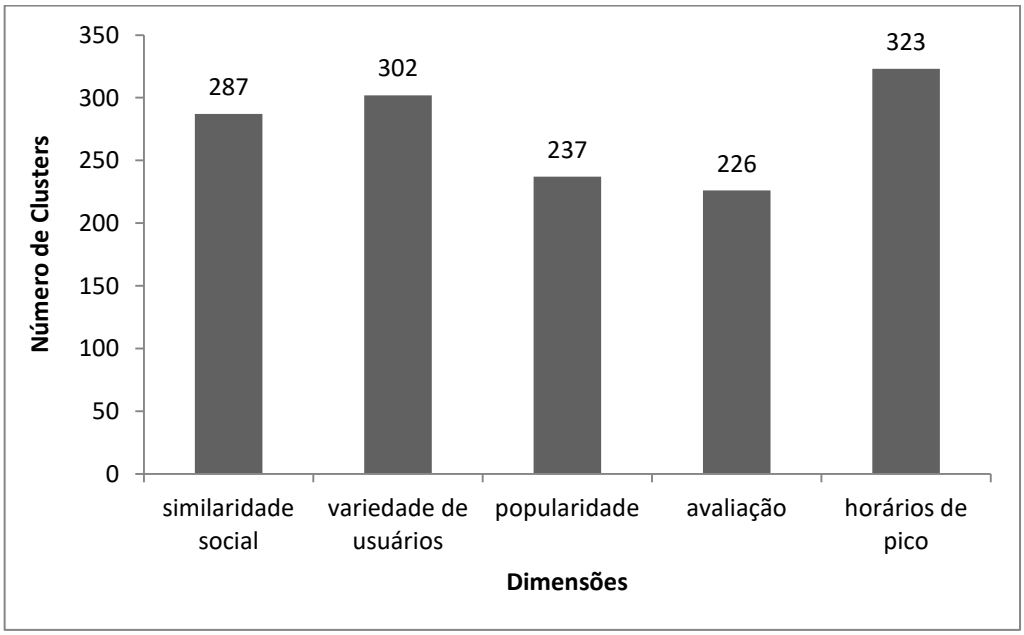
semelhantes entre si, a quantidade de *clusters* é uma medida mais relativa. Por exemplo, pode-se desejar uma menor quantidade para encontrar grandes regiões da cidade ou uma maior quantidade para encontrar atividades específicas de regiões. A Tabela 5.2 exibe a comparação dos coeficientes entre as dimensões e a Figura 5.2 apresenta a quantidade de *clusters* encontrados por dimensão. As subseções a seguir descrevem em mais detalhes a análise desta comparação.

Tabela 5.2 – Comparação entre as representações e o coeficiente resultante de coesão-separação

Modelo	*Coeficiente*
Similaridade Social (*baseline)*	41.25
Variedade de usuários	0.39
Popularidade social	0.63
Avaliação	0.19
Horários de Pico	0.36

Fonte: Rapacki (2017).

Figura 5.2 – Número de *clusters* descobertos para cada dimensão



Fonte: Rapacki (2017).

5.1.3.1 Similaridade Social

É possível observar que o modelo *baseline* de Similaridade Social (Cranshaw et al., 2012) obteve o melhor coeficiente de separação-coesão por uma grande margem. Isto indica que estabelecimentos que pertencem ao mesmo *cluster* possuem uma alta similaridade e os *clusters* são distintos um dos outros.

Esta grande diferença entre o modelo *baseline* e as dimensões propostas pode ser explicada por diversos motivos. Por exemplo, pode indicar a capacidade de Similaridade Social ser o melhor modo de encontrar regiões similares ou que o método da avaliação de coesão-separação beneficia este modelo por não ter características adicionadas a ele como os outros. Além disso, a dimensão possui um número médio de *clusters* quando comparado com os outros, o que não indica diretamente nenhum significado especial.

5.1.3.2 Variedade de Usuários e Horários de Pico

Por obterem resultados muito similares nas avaliações, ambas dimensões são apresentadas nesta mesma subseção. Seus coeficientes de coesão-separação apresentam valores médios em relação aos outros, mas a quantidade de *clusters* encontrados é a maior entre todas dimensões.

Uma possível explicação para isso é que estas dimensões são mais restritivas, analisando estabelecimentos próximos por seus horários de atividade, categoria e diversidade. Em outras palavras, são encontrados *clusters* menores, pois muitos locais geograficamente próximos não possuem a mesma diversidade ou distribuição de visitas, logo impossibilitando conectar regiões maiores.

5.1.3.3 Popularidade

O modelo de Popularidade possui a segunda menor quantidade de regiões e o melhor coeficiente entre as dimensões propostas por KANDOR. Deste modo, mesmo descobrindo regiões de tamanho maior, o coeficiente demonstra que os *clusters* encontrados demonstram características de coesão e separação quando comparados com as outras dimensões.

Analisando os padrões, pode-se supor que esta dimensão possui a maior capacidade de encontrar regiões semelhantes por seu caráter mais social, assim como o modelo *baseline* Similaridade Social.

5.1.3.4 Avaliação

De todos os resultados, a dimensão Avaliação apresenta o pior coeficiente de coesão-separação e a menor quantidade de *clusters*. Enquanto o número de regiões indica que essa

dimensão foi capaz de conectar um maior número de estabelecimentos, o coeficiente aponta uma fraca coesão e separação.

Este fenômeno pode ser explicado pelo fato da dimensão Avaliação ser uma característica mais geral de estabelecimentos e não existe uma relação direta entre estabelecimentos próximos possuírem a mesma avaliação. Assim, muitos estabelecimentos são introduzidos na mesma região com baixa similaridade, introduzindo ruído.

5.1.3.5 Análise Geral

Comparando a representação *baseline* de Similaridade Social e as representações propostas, é possível observar que o *baseline* obteve um resultado muito superior em relação ao coeficiente de coesão-separação. Entre as dimensões propostas, Popularidade gerou *clusters* com um melhor coeficiente, o que pode indicar uma melhor capacidade de dimensões com característica social encontrar regiões similares ou um viés do método de avaliação.

Em contrapartida, outros atributos como Avaliação, Variedade de Usuários e Horários de Pico encontram regiões com menor coeficiente de coesão-separação porque os locais possuem uma variedade maior de tipos. Por exemplo, uma pequena área da cidade pode conter restaurantes, padarias e lojas, que possuem diferentes diversidades e horários de pico.

O fato de que o modelo de Avaliação possui o menor coeficiente e o menor número de regiões pode potencialmente ser explicado pela dimensão não ser um bom indicador de similaridade. Apesar de ter encontrado regiões similares maiores, o coeficiente indica que os *clusters* não são coesos nem tão distintos entre si. Como todas dimensões usaram o mesmo parâmetro *m* para a quantidade de vizinhos mais próximos, essa diferença pode ser explicada pelo fato que alguns algoritmos usam dimensões que conectam regiões maiores e outros são mais específicos para regiões locais menores.

5.2 Avaliação com Usuários através de Interface Visual

KANDOR propõe um método para gerar diferentes formas de representação de *clusters* e sugere cinco dimensões para verificar a diferença das regiões da cidade encontradas por elas. O objetivo deste experimento é avaliar com usuários o potencial de cada dimensão para encontrar regiões relevantes da cidade através de suas percepções e conhecimento da região. De modo a avaliar a contribuição das dimensões propostas, foi utilizada uma

metodologia de pesquisa centrada em um questionário *online* com pessoas que moram ou já moraram em Porto Alegre. Estes usuários tinham como tarefa alternar entre o questionário e a página *web* para analisar a cidade a partir de cada dimensão e responder perguntas sobre suas percepções. As seções a seguir descrevem o protocolo estabelecido para o experimento, a demografia, as hipóteses e os resultados encontrados.

5.2.1 Protocolo

41 pessoas participaram do experimento como voluntárias, recrutadas por um anúncio em um *e-mail* e por uma publicação no Facebook entre 25 de Novembro e 18 de Dezembro, 2017. Os convites para o experimento foram enviados em diferentes momentos para diferentes públicos, com o objetivo de, primeiro, coletar feedback com os primeiros usuários para aperfeiçoar o questionário e, depois, avaliar a semântica dos agrupamentos gerados pela Interface de visualização desenvolvida neste trabalho.

5.2.1.1 Experimento Piloto

Primeiramente, dois alunos do grupo de pesquisa "Sistemas de Informação" foram convidados para participar através de um *e-mail* direto com uma breve explicação do trabalho e o *link*[9] para o questionário. Uma semana depois, foi enviado um e-mail similar para o restante dos alunos do grupo de pesquisa e outros mestrandos do Instituto de Informática da UFRGS. Dois dias depois, foi publicado no Facebook um convite para uma audiência mais geral com uma breve descrição e o mesmo *link* para o questionário.

Este processo foi separado dessa forma para aplicar inicialmente o questionário em grupos menores, mais familiarizados com os termos utilizados no trabalho e possivelmente até com o trabalho em si. Por exemplo, uma parte dos alunos do grupo de pesquisa assistiu a uma apresentação que mostrava uma visão geral do trabalho e por isso tinham mais conhecimento da proposta. Além disso, os primeiros grupos de acadêmicos de Informática estão mais acostumados com conceitos de agrupamento, análise de dados e navegação em páginas *web* e portanto tendem a ter mais facilidade para executar as tarefas. Após a execução da pesquisa com os dois primeiros usuários, foi possível coletar opiniões sobre questões de usabilidade da ferramenta e tornar as tarefas e perguntas mais claras no questionário. Assim, o

[9] https://goo.gl/forms/YYGJL7Gd196KKyQh1

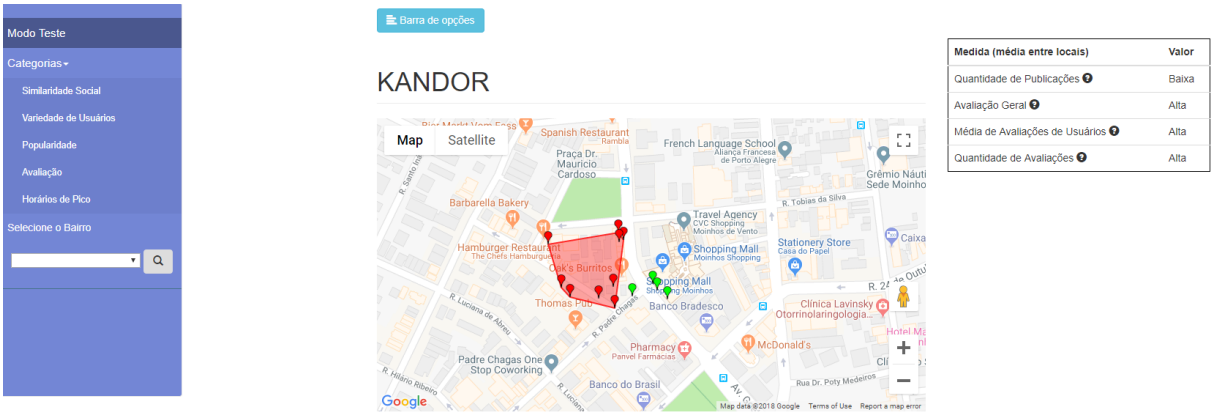
protocolo se manteve o mesmo entre todos usuários, com o texto descritivo aprimorado para ficar mais claro e fácil após os dois primeiros grupos (dois alunos e o resto do grupo de pesquisa).

5.2.1.2 Experimento

O questionário (que está no Anexo A) enviado consiste de quatro partes: Informações pessoais; um modo teste para avaliar a influência da ferramenta no resultado; um tutorial explicando as tarefas a serem executadas; e três conjuntos similares de tarefas para bairros diferentes.

O modo teste foi introduzido para demonstrar que a usabilidade da interface não é um fator influenciável no desempenhoda dimensão, solicitando ao usuário a execução de uma tarefa bem simples com a ferramenta. Desta maneira, pode-se comparar o desempenho dos usuários no modo teste e suas respostas e verificar se há influência na interação com a ferramenta. Para isso, tarefas triviais baseadas nas apresentadas no experimento principal são solicitadas, testando isoladamente a capacidade de compreender as instruções e interagir com a interface. São apresentadas duas regiões no mapa em um bairro específico, cada uma com pontos de cores distintas e específicas para cada região. Em seguida, o usuário é instruído a navegar pela interface e responder quantas regiões e quantos locais em cada região foram encontrados. Finalmente, uma última pergunta solicita ao usuário a analisar a tabela de medidas e responder de forma livre quais são as principais diferenças nas medidas das duas regiões. Na sequência, o usuário recebeu um texto (Tutorial que está no Anexo A) explicando os conceitos básicos da interface e as dimensões propostas. As tarefas do experimento principal são divididas em três conjuntos, cada um para um tipo diferente de bairro.

Figura 5.3 – Página exibindo o modo teste com uma região selecionada



Fonte: Rapacki (2017).

A escolha do bairro para cada uma das seções é definida pelo contexto: a primeira é o bairro de moradia do usuário, a segunda é o bairro onde o usuário passa suas horas úteis do dia trabalhando ou estudando, chamada de rotina, e a última é o bairro que o usuário frequenta por lazer. O motivo para definição desses três contextos é avaliar se os horários que o usuário frequenta o bairro ou o tipo de atividade pode favorecer uma ou outra dimensão.

Para cada bairro, o mesmo conjunto de tarefas orienta o usuário a analisar as regiões geradas por cada dimensão em relação ao bairro em questão. Como visto na seção 4.2, o menu da esquerda da Interface auxilia o usuário nessa tarefa apresentando as cinco dimensões (na Interface exibido como categoria) disponíveis para gerar os agrupamentos da cidade. O usuário é instruído a executar cada tarefa para uma determinada dimensão na página *web* e responder a pergunta correspondente no questionário, alternando entre um e outro. Após a execução da tarefa, o usuário deve assinalar uma escala Likert[10] de a 1 a 5 sobre duas afirmações relativas a relevância da dimensão. A primeira avalia o quanto as regiões geradas são determinadas pela dimensão, ou seja, se os locais em cada região são semelhantes em relação ao que a dimensão propõe. A segunda explora as regiões que ultrapassam os limites do bairro, avaliando se são áreas que estão se propagando ou se integrando com outros bairros.

Adicionalmente, algumas variáveis foram coletadas no questionário demográfico e no modo teste, ajudando a definir o perfil do usuário. Os dados do usuário consistem em idade, escolaridade, profissão, tempo de moradia na cidade, frequência de uso de redes sociais e

[10] Uma das metodologias mais populares para realizar pesquisas de opinião. As questões apresentam uma afirmação auto-descritiva e oferecem como opção de resposta uma escala de pontos que contemplam extremos. Deste modo, é possível avaliar diferentes níveis de intensidade de opinião em relação a pergunta.

frequência nos bairros. A performance do usuário no modo teste é a soma de acertos para as tarefas, avaliando se a interface influenciou a performance do usuário no questionário.

As variáveis independentes deste experimento são o tipo de bairro sendo analisado (moradia, rotina e lazer) e a dimensão utilizada para gerar os agrupamentos. Enquanto isso, as variáveis dependentes são o desempenho da dimensão (o quanto os locais são percebidos como semelhantes dentro de uma região) e a capacidade de revelar propagações através dos limites oficiais do bairro.

Finalmente, como os voluntários participaram por livre e espontânea vontade e o questionário era relativamente longo e complexo, assume-se que todas pessoas que responderam todo o questionário seguiram os passos na ordem proposta e o mais honestamente possível.

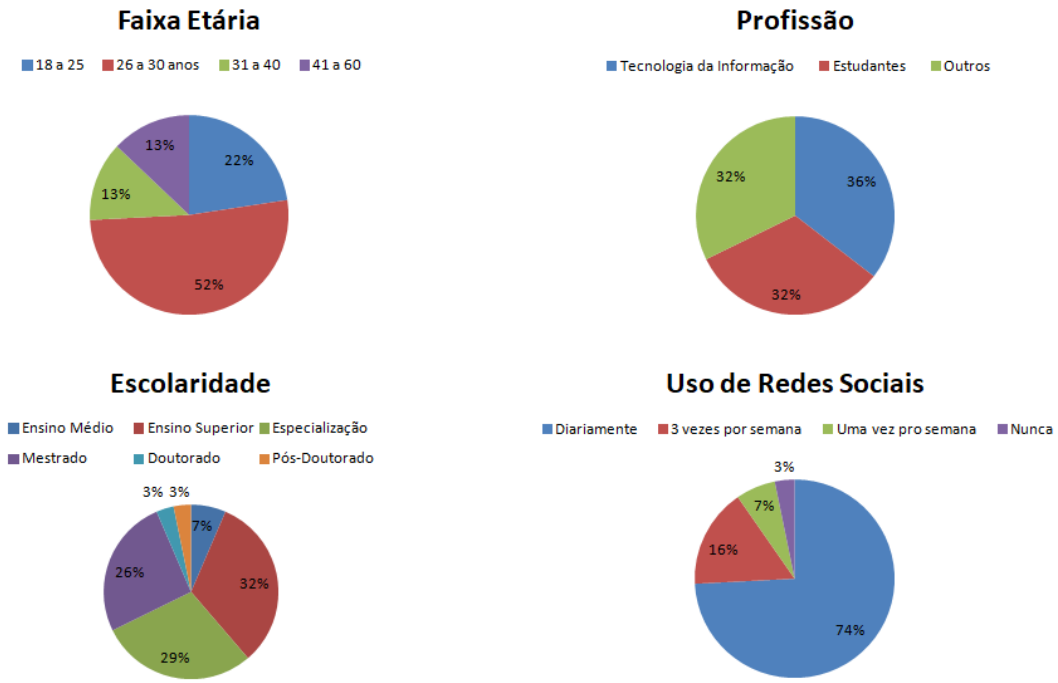
5.2.2 Demografia

Essa seção descreve a demografia dos participantes do experimento e tem como objetivo mapear as suas características e cruzá-las com as respostas para buscar possíveis padrões. Foram realizadas as seguintes perguntas:

1. Você mora em Porto Alegre?
2. [Se respondeu NÃO pra pergunta acima] Já morou alguma vez em Porto Alegre? Por quanto tempo?
3. Qual sua idade?
4. Qual sua profissão?
5. Qual seu nível de escolaridade?
6. Você possui um smartphone (celular inteligente com conexão com a internet)?
7. Se respondeu sim pra pergunta acima, com qual frequência você utiliza redes sociais como Facebook, Instagram ou Foursquare?

No total, 41 pessoas responderam o questionário enviado, porém 10 desistiram no meio e submeterem o questionário incompleto, portanto somente 31 respostas são válidas. Destas 31 pessoas, 28 moram atualmente em Porto Alegre e 3 não. Para as três pessoas que não moram em Porto Alegre, o período que elas moraram em Porto Alegre foi categorizado em três intervalos para diferenciação: menos de 5 anos, entre 5 a 10 anos e mais de 10 anos. Essas três pessoas que não moram em Porto Alegre assinalaram um período diferente cada uma. A Figura 5.4 abaixo apresenta os gráficos com o percentual das principais características como idade, profissão, escolaridade e uso de redes sociais.

Figura 5.4 – Gráficos exibindo a demografia dos usuários



Fonte: Rapacki (2017).

Pouco mais da metade das pessoas possuem entre 26 e 30 anos, 7 possuem entre 18 e 25 anos e as idades de 31 a 40 e 41 a 60 possuem 4 usuários cada. A média de idade é de aproximadamente 30 anos e o desvio padrão é 8.88. As profissões dos indivíduos varia entre funcionários de Tecnologia da Informação como desenvolvedores e analistas (11), estudantes na sua maioria mas não exclusivamente de Informática (10) e outras mais gerais (10), como médico, contador, publicitário, etc. Em sua maioria, os entrevistados possuem escolaridade acima do Ensino Médio: 10 possuem Ensino Superior, 9 possuem alguma especialização, 8 possuem Mestrado, um possui Doutorado e um Pós-Doutorado. Os dois restantes possuem Ensino Médio completo.

Em relação a *smartphones* e redes sociais, todas pessoas possuem algum tipo de *smartphone* e a grande maioria utiliza redes sociais diariamente. O restante se divide em 5 pessoas que usam aproximadamente 3 vezes por semana, 2 que usam uma vez por semana e uma que não usa.

5.2.3 Hipóteses

Neste experimento, é formulada uma hipótese a ser validada ou não pelo questionário. A hipótese contempla o objetivo do KANDOR e define que as diferentes dimensões (similaridade social, variedade de usuários, etc.) fornecem visões diferentes da cidade, adicionando assim uma gama maior de informações para a sua compreensão. Assim, a causa da avaliação da dimensão é efeito do contexto em que o bairro é analisado. Por exemplo, certas dimensões podem possuir uma avaliação melhor quando o usuário frequenta o bairro por lazer e outras quando residem no bairro. No experimento, isso pode ser observado analisando as notas dos usuários para cada dimensão e cada tipo de bairro. Se houver uma variação das notas da mesma dimensão para diferentes contextos, há um indício que a hipótese é verdadeira. Porém, é importante ressaltar que nem todas dimensões precisam obter notas altas, uma vez que este trabalho busca somente explorar o benefício da utilização de outros tipos de informações para entender as divisões da cidade.

Para comprovar a hipótese, assume-se a hipótese nula de que a avaliação das dimensões não é determinada pelo contexto. Isto significa que, ou a causa das avaliações é variação randômica ou é determinada somente pelo desempenho da dimensão em si, e não do contexto. Se a última parte for verdadeira, significa que algumas dimensões teriam avaliação consistentemente superior a outras independente do contexto.

5.2.4 Resultados

Esta seção descreve os resultados do experimento, analisando os dados brutos e também comparando diferentes variáveis para analisar suas correlações. Assim, o objetivo é descobrir se as hipóteses formuladas são válidas e quais foram as dimensões que obtiveram resultados positivos e quais foram indiferentes.

As perguntas feitas no questionário sobre as dimensões representam duas variáveis a serem analisadas: o seu desempenho (o quanto os locais se assemelham dentro de uma região) e a capacidade de revelar propagações através dos limites oficiais do bairro. São avaliados também os resultados do modo teste, de modo a observar se a resposta do usuário influencia em suas notas para as dimensões.

5.2.4.1 Modo Teste

Antes de fornecer as instruções e as tarefas para o usuário, um modo teste foi incluído para avaliar o entendimento do usuário no uso da ferramenta e o seu desempenho na execução das tarefas solicitadas. Assim, cada resposta foi considerada um acerto se for exatamente a esperada e o número de acertos é somado para cada usuário. O modo teste exibe uma visualização da cidade com somente duas regiões, demarcadas por cores diferentes. As primeiras três perguntas feitas no questionário testam se o usuário é capaz de identificar quantas regiões foram encontradas e quantos estabelecimentos pertencentem a cada região. A última pergunta pede para o usuário analisar as diferenças das duas regiões em relação a duas tabelas de métricas (quantidade de *check-ins* e quantidade de usuários únicos).

Analisando os resultados, observou-se que 20 pessoas responderam corretamente todas as quatro perguntas e 7 pessoas acertaram três, normalmente cometendo o erro por uma margem pequena. As 4 pessoas restantes podem ter encontrado dificuldades na ferramenta ou na compreensão da tarefa, o que pode afetar suas respostas no restante do questionário. Na subseção 5.2.4.2, a correlação entre as variáveis será analisada e assim será possível observar se desempenho no modo teste influenciou no desempenhoda dimensão ou não.

5.2.4.2 Pré-Análise: Dados Brutos

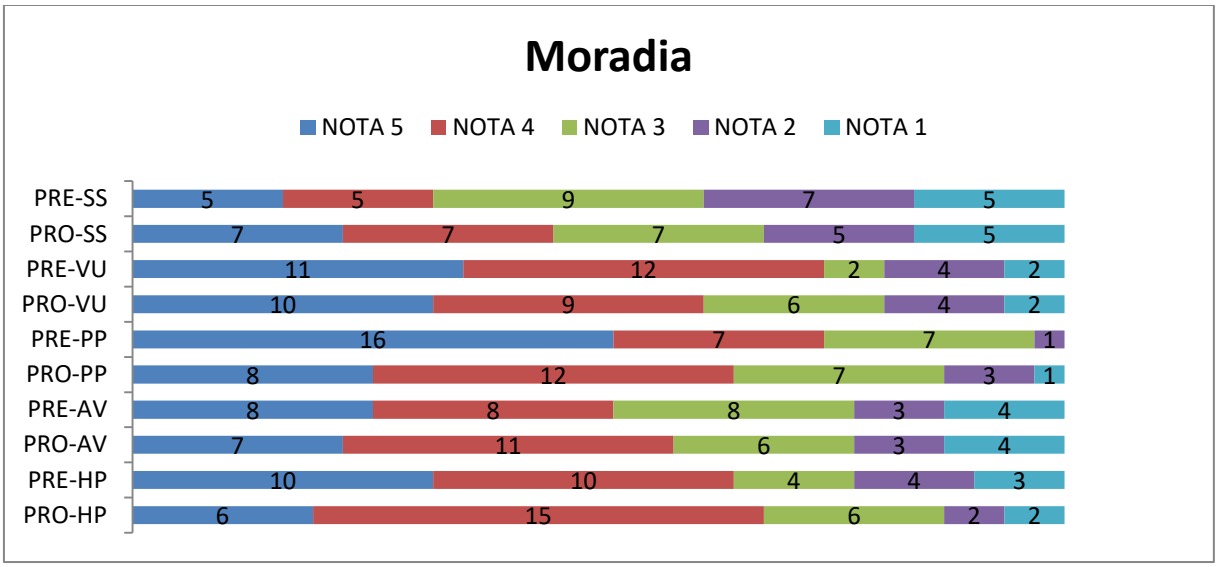
Para analisar as avaliações obtidas para cada dimensão, as duas métricas apresentadas na subseção 5.2.1 foram utilizadas: desempenho e capacidade de revelar propagações. Para isto, três conjuntos de tarefas guiam o usuário através de 10 perguntas cada (2 por dimensão), repetindo para o bairro de moradia, de rotina e de lazer. Na primeira subseção 5.2.4.1.1, são apresentados gráficos com a quantidade de notas atribuídas pelos usuários para cada tarefa, agrupadas por tipo de bairro.

As legendas na esquerda dos gráficos representam qual a dimensão e contexto foram avaliados: DES significa o desempenho da dimensão e PRO a propagação do bairro. Assim, pode-se comparar o desempenhoe propagação de cada dimensão para o mesmo tipo de bairro. Em seguida, a próxima subseção apresenta as mesmas notas porém agrupadas por dimensão, mostrando como a mesma dimensão é avaliada para os diferentes tipos de bairro.

5.2.4.2.1 Notas Agrupadas por Tipo de Bairro

Notas agrupadas por tipo de bairro permitem analisar a correlação entre o desempenhoe propagação das dimensões com o contexto da região, por exemplo por qual motivo o usuário frequenta no bairro. O gráfico da Figura 5.5 apresenta o número de avaliações obtidas por cada dimensão para o bairro de moradia. Assim como o nome das notas foi abreviado para PRE e PRO, os nomes das dimensões também foram abreviados: SS (similaridade social), VU (variedade de usuários), PP (popularidade), AV (avaliação) e HP (horários de pico).

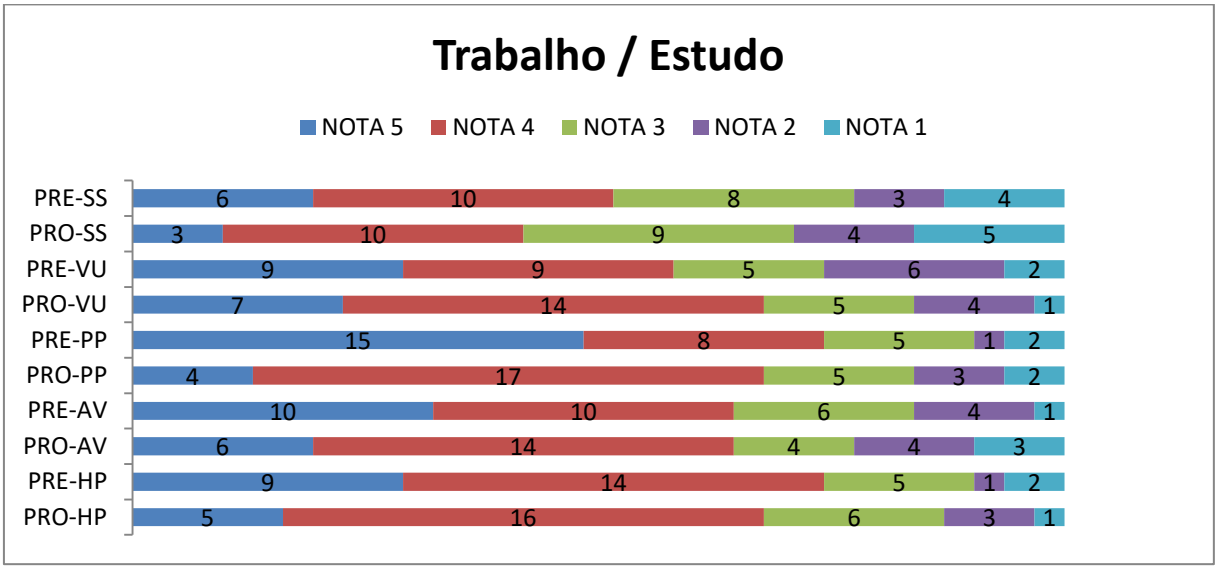
Figura 5.5 – Avaliações de desempenho e propagação por dimensão para Moradia



Fonte: Rapacki (2017).

Primeiramente, é visível que a dimensão Similaridade Social obteve a pior avaliação de todas, possuindo menos da metade de votos com notas altas (5 ou 4). Alguns comentários dos usuários elaboram essa questão, por exemplo: “No meu bairro, o sistema não apresentou muitos pontos (e não achou nenhuma região) para a dimensão similaridade social”. A dimensão Avaliação aparece em segundo lugar com notas médias, o que pode significar que não possui uma contribuição significativa para este tipo de bairro. As três dimensão restantes apresentaram notas boas tanto no desempenho quanto na propagação com a soma de notas altas entre 19 e 23. É possível observar também que comparando as duas medidas para a mesma dimensão, Popularidade e Variedade de Usuários tiveram notas maiores para desempenho e as três restantes foram levemente melhores para propagação.

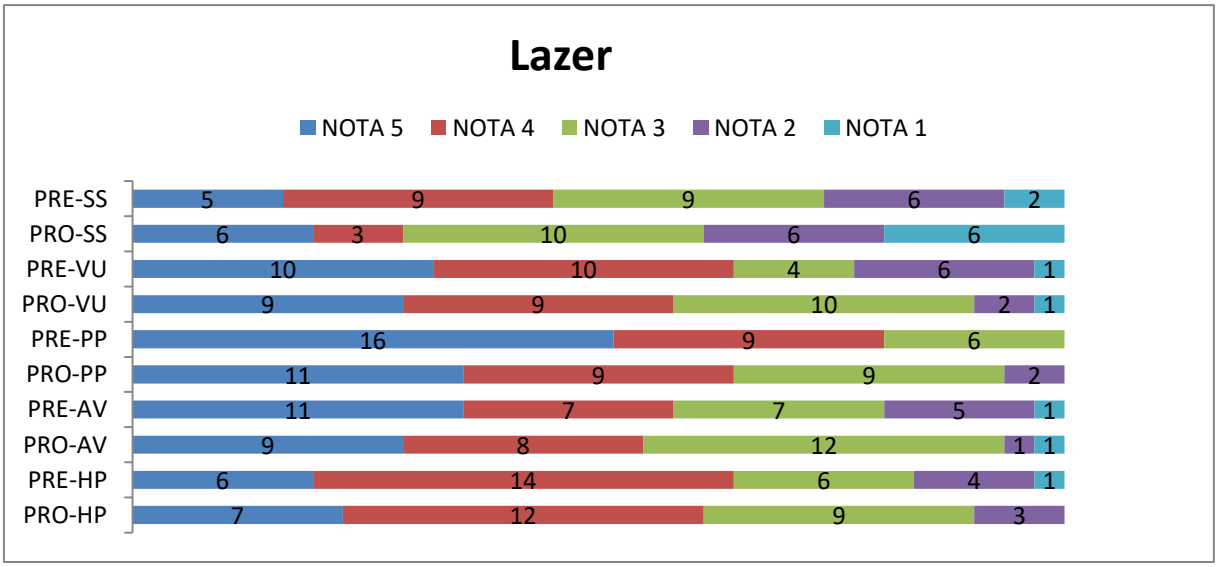
Figura 5.6 – Avaliações de desempenho e propagação por dimensão para Trabalho/ Estudo



Fonte: Rapacki (2017).

Para o bairro onde os usuários trabalham ou estudam, os resultados exibidos na Figura 5.6 foram similares mas algumas diferenças foram encontradas. Similaridade Social obteve um desempenho um pouco melhor, mas continua sendo a pior das cinco dimensões. Em seguida, Avaliação e Variedade de Usuários alcançam valores médios entre 18 e 21 e enfim Horários de Pico e Popularidade se destacam com valores entre 20 e 23. Nesses resultados, Popularidade desempenhou melhor em desempenho do que propagação, Avaliação obteve resultados iguais e as demais foram melhores em propagação.

Figura 5.7 – Avaliações de desempenho e propagação por dimensão para Lazer



Fonte: Rapacki (2017).

Finalmente, quando o bairro analisado é um local em que os usuários frequentam por lazer, fica claro na Figura 5.7 a dimensão perdedora e vencedora neste contexto. Novamente, Similaridade Social fica com a pior nota e desta vez Popularidade se destaca sozinha como a melhor dimensão com 25 notas altas para desempenho. As outras três ficam com valores muito próximos entre 18 e 20 notas altas. Além disso, todas as dimensões obtiveram notas maiores no seu desempenho do que na sua propagação.

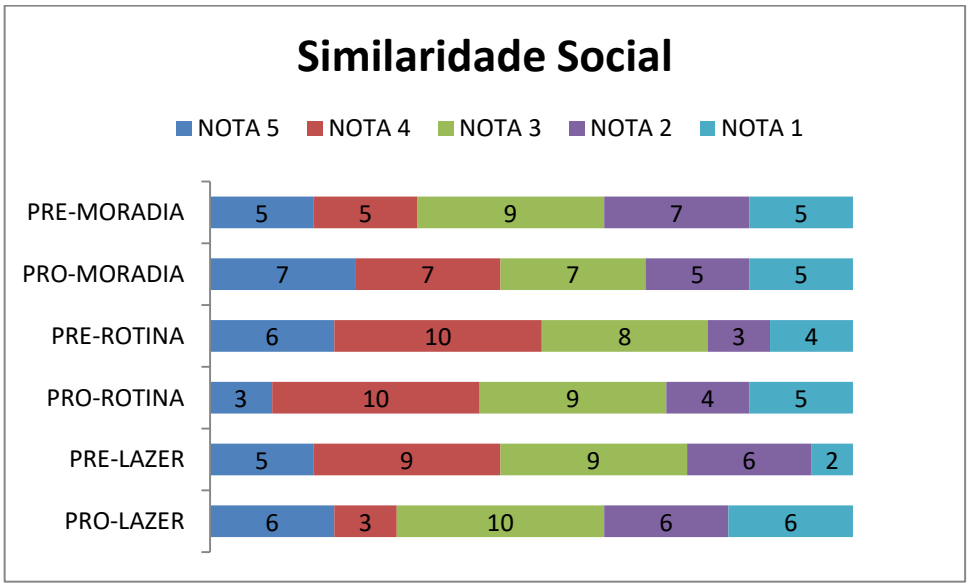
Observando as respostas descritivas dos usuários para cada bairro, pode-se detectar algumas tendências. Por exemplo, em geral a dimensão Similaridade Social obteve notas ruins porque gerou poucos pontos, poucas regiões e poucas ou nenhuma região que ultrapassem os limites do bairro. Um usuário comentou nas respostas: “Poucos dados disponíveis para a dimensão Similaridade Social. O restante tinha bastante informação”. Ao comparar o número de regiões e pontos gerados, realmente pode-se notar que Similaridade Social não obteve o mesmo desempenho das outras. Uma das possíveis causas para isto é que como o valor *m* escolhido para o *clustering* foi o melhor em média para todas dimensões, esse pode não ter sido o melhor valor para similaridade social. Isso significa que com esse valor não foi possível conectar muitos locais próximos em Porto Alegre e por isso as regiões foram escassas.

Além disso, muitos usuários comentaram que alguns bairros possuem poucos pontos dispersos e que a grande maioria das regiões possui quantidade baixa de publicações, que pode prejudicar o desempenho de algumas dimensões principalmente Similaridade Social e Variedade de Usuários. Entretanto, houve também opiniões interessantes reforçando a importância de possuir diferentes dimensões e diferentes tipos de bairro sendo analisados: um usuário afirma “Não foi possível avaliar os valores (da tabela), quanto às dimensões achei relevante”. Outro comenta: “Sobre este bairro (lazer) respondi com mais firmeza sobre sua propagação e integração com outros, pois percebo mais esse movimento quando frequento o bairro”.

5.2.4.2.2 Notas Agrupadas por Dimensão

Notas agrupadas por dimensão tem como objetivo analisar o desempenho e a propagação da mesma dimensão para todos os tipos de bairro. Desta maneira, pode-se observar se existem variações nas percepções dos usuários para algum contexto diferente de bairro.

Figura 5.8 – Avaliações de desempenho e propagação por tipo de bairro para Similaridade Social

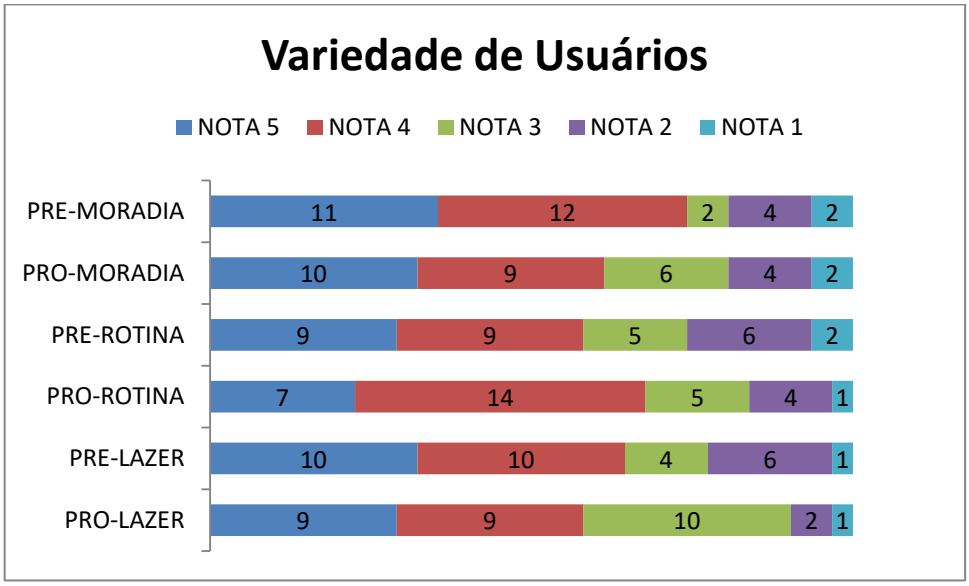


Fonte: Rapacki (2017).

A Figura 5.8 exibe as performances e propagações da dimensão Similaridade Social para cada tipo de bairro. Pode-se observar que enquanto o desempenho em bairros de moradia possui menos notas altas que em outros bairros, a propagação teve um menor valor em bairros de lazer. Isto pode fornecer indícios de que esta dimensão encontra regiões mais semelhantes sob a perspectiva de atividades sociais (rotina e lazer). Além disso, como a dimensão obteve notas piores de propagação para o contexto de lazer, há a possibilidade de que o público dos locais de entretenimento costume se restringir ao mesmo bairro.

Para a dimensão Diversidade de Usuários, é possível observar na Figura 5.9 que o melhor desempenho é obtida em bairros de moradia, enquanto a melhor propagação é encontrada em bairros de rotina. Adicionalmente, a pior nota obtida de desempenho é em bairros de rotina e, para a medida de propagação, bairros de moradia e lazer virtualmente empatam. Desta maneira, pode-se especular que a dimensão consegue encontrar regiões com diversidade de usuário mais semelhante em bairros residenciais e que bairros de rotina possuem um maior potencial de se conectar com outros bairros com uma diversidade similar.

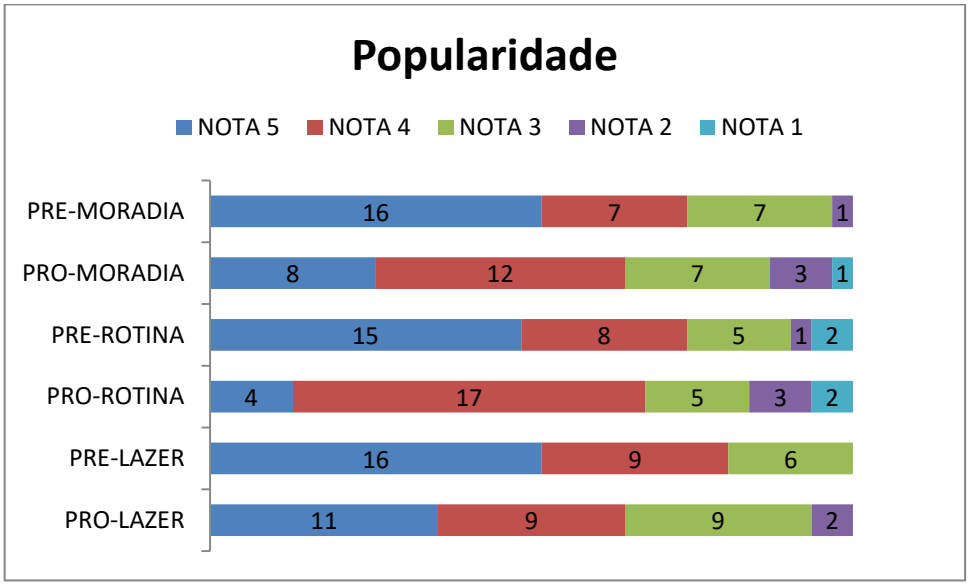
Figura 5.9 – Avaliações de desempenho e propagação por tipo de bairro para Variedade de Usuários



Fonte: Rapacki (2017).

A Figura 5.10 exibe que para Popularidade, bairros de lazer obtém o maior desempenho entre os tipos de bairro. Logo, uma possível explicação é que os estabelecimentos de regiões em bairros de lazer possuem uma influência social semelhante. Isso representa a tendência que existe em cidades onde algumas regiões específicas concentram estabelecimentos que as pessoas frequentam por lazer. Além disso, para todos tipos de bairro, poucas notas baixas (1 e 2) foram atribuídas para esta métrica o que representa o potencial de Popularidade para qualquer contexto. Para propagação, os três tipo de bairro alcançam praticamente com a mesma nota.

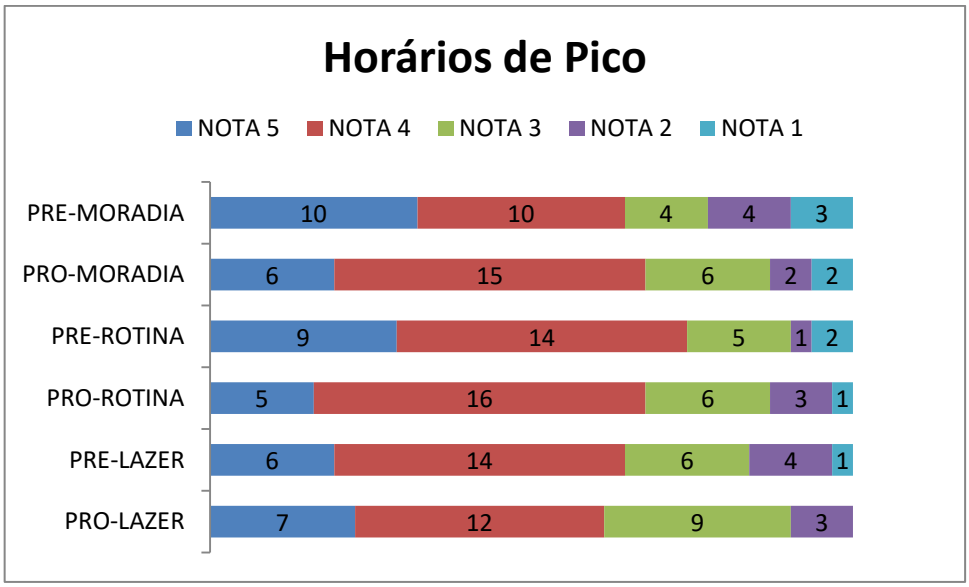
Figura 5.10 – Avaliações de desempenho e propagação por tipo de bairro para Popularidade



Fonte: Rapacki (2017).

A dimensão Horários de Pico alcança notas médias no geral, com bairros de rotina prevalecendo no desempenho, como mostrado na Figura 5.11. Essa vantagem pode ser explicada pelo fato de que bairros de rotina (estudo ou trabalho) possuem uma atividade mais restrita a horários úteis do dia e por isso consegue encontrar regiões mais semelhantes. Em geral, essa dimensão aparenta atingir valores médios de desempenho e propagação não importando qual o tipo de bairro escolhido.

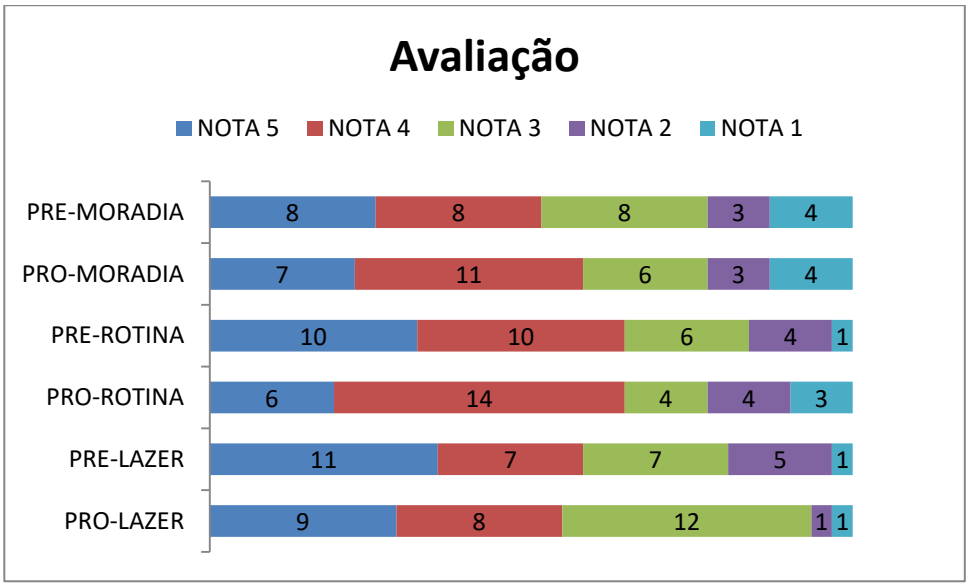
Figura 5.11 – Avaliações de desempenho e propagação por tipo de bairro para Horários de Pico



Fonte: Rapacki (2017).

Analisando as dimensões Variedade de Usuários, Popularidade e Horários de Pico, é possível observar que os resultados obtidos de performance e propagação para cada tipo de bairro ficam muito próximos. Isso pode indicar à primeira vista que não existe uma dependência forte entre a dimensão e o contexto analisado.

Figura 5.12 – Avaliações de desempenho e propagação por tipo de bairro para Avaliação



Fonte: Rapacki (2017).

Por outro lado, há um indício de que a dimensão Avaliação, exibida na Figura 6.8, possui desempenho e propagação melhores para o contexto de atividades rotineiras, o que pode indicar que os usuários conhecem mais a qualidade do bairro e arredores dado que costumam frequentar com mais frequência. É interessante observar também que a propagação no contexto de lazer teve 29 de 31 votos nota 3 ou superior, o que pode representar que avaliação é um conceito bastante abrangente, conseguindo conectar locais de diferentes bairros.

5.2.4.3 Análise

Com o objetivo de avaliar a relação entre as variáveis independentes e dependentes, esta seção apresenta alguns testes estatísticos que foram realizados sobre os resultados do questionário. Primeiramente, são comparados os dados do perfil de usuário como idade e profissão com o resultado do modo teste, buscando evidenciar alguma relação com o desempenho e a compreensão da ferramenta.

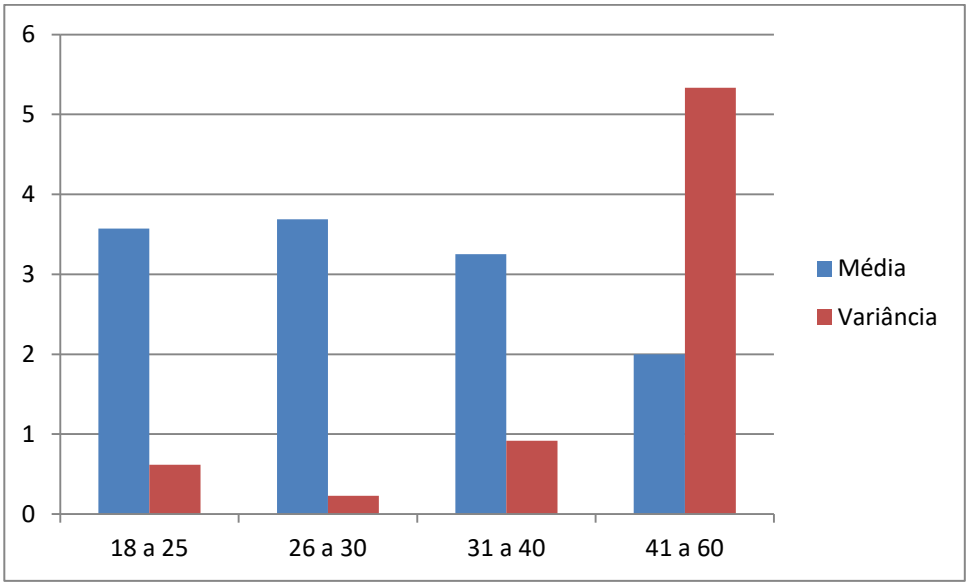
Em seguida, comparam-se as duas variáveis independentes, dimensão e tipo de bairro, com as respostas do questionário para examinar a correlação entre elas. As correlações entre variáveis são analisadas com diferentes tipos de Análise de Variância (ANOVA), o de um fator e de dois fatores. ANOVA é um método para testar a média de três ou mais médias populacionais, analisando as variâncias das amostras e indicando se os resultados são estatísticamente significantes. Um fator na ANOVA é uma característica que permite diferenciar as populações uma das outras. Além disso, pode-se usar ANOVA com replicação quando existe mais de um grupo sendo analisado. Assim, nas subseções seguintes foi utilizado ANOVA de um fator e de dois fatores com replicação conforme o número de características que se deseja analisar a correlação.

5.2.4.3.1 Idade X Modo Teste

Para a primeira análise entre as variáveis do perfil de usuário e o modo teste, foi escolhida a idade dos usuários para observação. A população foi separada em quatro intervalos de idade e o número de acertos para as quatro perguntas foi atribuído para cada pessoa dentro de seu intervalo. Utilizando o teste ANOVA de um fator, é possível determinar que há uma diferença estatisticamente significante entre as médias dos resultados para cada idade ($p < 0.03$). Calculando a correlação de Pearson entre as variáveis, observa-se uma correlação média negativa entre elas, ou seja pessoas nos grupos mais jovens tiveram mais facilidade no uso e compreensão da ferramenta.

A Figura 6.9 exibe a média e variância dos acertos por faixa etária dos entrevistados. Pode-se observar uma variância maior nos acertos dos entrevistados na faixa de 41 a 60 anos enquanto nas outras faixas a variância permaneceu baixa. Enquanto isso, a média de acertos foi maior para o intervalo de 26 a 30 anos e entre 18 a 25 e 31 a 40 a média foi similar.

Figura 5.13 – Média e Variância de acertos por faixa etária



Fonte: Rapacki (2017).

5.2.4.3.2 Escolaridade e Profissão X Modo Teste

Ao examinar a interação entre as variáveis de escolaridade e profissão com o desempenho no modo teste, foi utilizado novamente o teste ANOVA de um fator para cada uma das variáveis independentes. Entretanto, desta vez não foi possível encontrar uma diferença estatísticamente significante entre os diferentes grupos de cada variável, o que comprova que não há uma correlação com o desempenho com a ferramenta.

5.2.4.3.3 Tipo de Bairro X Dimensão

Após determinar que somente a idade influencia no desempenho com a interface, são analisadas as variáveis independentes do tipo de bairro e dimensão com as avaliações. Para isso, foi utilizado um teste ANOVA de dois fatores com replicação, um para desempenho e outro para propagação. Os resultados do teste demonstraram que a variável dimensão isoladamente consegue encontrar diferenças de médias estatisticamente significantes, porém o mesmo resultado não foi alcançado com a combinação de dimensão x tipo de bairro ou somente tipo de bairro. Isto significa que para as amostras deste trabalho, consegue-se

concluir somente que a dimensão utilizada gera diferentes visões da cidade, mas que não necessariamente são dependentes do contexto do bairro.

Em seguida, foram removidas as respostas dos quatro usuários que atingiram menos de 3 acertos no modo teste e foi aplicado novamente o teste ANOVA nesta nova amostra. Entretanto, novamente obteve-se o mesmo resultado de que somente a dimensão é um fator determinante para os resultados de desempenho e propagação.

6 CONCLUSÃO

Nesta dissertação, endereça-se o problema de identificar regiões semelhantes de uma cidade para compreender melhor a sua estrutura e dinâmica. É proposto o método KANDOR para gerar formas de representação de *clusters* a partir de diferentes tipos de dados de redes sociais, por exemplo: co-locação, mobilidade de usuário, categoria de atividade, influência social, etc. Esse método é composto por cinco dimensões que representam diferentes características de – similaridade social, diversidade de usuários, popularidade, avaliação e horários de pico. Para avaliar o método apresentado, realizaram-se experimentos de calibragem para o algoritmo de *clustering* espectral e uma pesquisa quantitativa com residentes de Porto Alegre.

Através de uma análise estatística dos *clusters* gerados, é possível observar que as dimensões geram regiões de tamanhos e graus de semelhança diferentes entre estabelecimentos na mesma região. A pesquisa com usuários também exibiu a capacidade das dimensões de gerar regiões com desempenho e propagação diferentes de acordo com o contexto analisado, por exemplo qual tipo de atividade o usuário realiza no bairro. Os resultados demonstram que informações de diferentes redes sociais e tipos podem ser agregadas de modo a fornecer visões diferentes da cidade.

Esse potencial permite que aplicações e usuários diferentes usufruam dos benefícios de visões específicas, como planejamento urbano, recomendação de locais, caracterização urbana, etc. Além disso, o método demonstra que novas formas de representação de *clusters* podem ser elaborados a partir de outras informações disponíveis em redes sociais, oferecendo inúmeras possibilidades de extensão de acordo com o objetivo. Como resultado dessa dissertação, foram publicados os seguintes artigos:

- *KANDOR – Knowledge Analysis of Neighborhood Dynamics and Online Relationships*; 2017 IEEE *41st Annual Computer Software and Applications Conference* (COMPSAC 2017); Qualis A2
- Utilizando Características Semânticas e Espaço-Temporais de Dados de *Social Media* para Descoberta de Conhecimento em *Smart Cities;* Workshop de Teses e Dissertações em Banco de Dados (WTDBD) no Simpósio Brasileiro de Banco de Dados (SBBD) 2016.

Uma das limitações deste trabalho é que o uso de redes sociais para representar atividades de usuário tende a excluir usuários que não são ativos nas redes sociais. Além

disso, apesar de descobrir a influência das dimensões nos *clusters* gerados, as análises foram inconclusivas para o uso do contexto do bairro.

Como trabalhos futuros, é possível realizar pesquisas mais extensivas de modo a visualizar melhor as diferenças entre as dimensões propostas. Além disso, uma etapa de pesquisa qualitativa possibilitaria a descoberta de informações mais intuitivas sobre o método. Outros benefícios também podem ser encontrados em tipos de informação não extraídos por KANDOR, como por exemplo análise de sentimento com os conteúdos e comentários das publicações do Instagram. Adicionalmente, as representações de *clusters* foram construídos com a mesma base (similaridade social) para aproximar a comparação entre eles, mas novas possibilidades podem ser encontradas ao utilizar representações restritivas a características de uma única dimensão.

REFERÊNCIAS

ALIGULIYEV, R. M. Performance evaluation of density-based clustering methods. **Information Sciences: an International Journal**, vol. 179, No. 20, 2009, p. 3583-3602.

BALDUINI, M.; BOCCONI, S.; BOZZON, A.; DELLA VALLE, E.; HUANG, Y.; OOSTERMAN, T.; PALPANAS, T.; TSYTSARAU, M. A case study of active, continuous and predictive social media analytics for smart city, **Proc. of the Fifth International Conference on Semantics for Smarter Cities (S4SC)**, vol. 1280, 2014, p. 31-46.

CRANSHAW, J.; SCHWARTZ, R.; HONG, J.; SADEH, N. The livehoods project: utilizing social media to understand the dynamics of a city, **Proc. of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)**, June 4-7, 2012, p. 58-65.

CRANSHAW, J.; TOCH, J.; HONG, J.; KITTUR, A.; SADEH, N. Bridging the gap between physical location and online social networks, **UbiComp'10 Proc. of the 12th ACM international conference on Ubiquitous Computing**, 2010, p. 119-128.

FRIAS-MARTINEZ, V.; FRIAS-MARTINEZ, E. Spectral clustering for sensing urban land use using Twitter activity, **Engineering, Applications of Artificial Intelligence**, vol. 35, October 2014, p. 237-245.

HASAN, S.; ZHAN, X.; UKKUSURI, S. V. Understanding urban activity and mobility patterns using large-scale location-based data from online social media, **UrbComp '13, Proc. of the 2nd ACM SIGKDD International Workshop on Urban Computing**, Article No 6, August 11-11, 2013, p. 1-8.

NOULAS, A.; SCELLATO, S.; MASCOLO, C.; PONTIL, M. An empirical study of geographic user activity patterns in Foursquare, **Proc. of the Fifth International Conference on Weblogs and Social Media (ICWSM '11)**, 2011, p. 570-573.

SILVA, T. H.; DE MELO, P. O. S. V.; ALMEIDA, J. M.; SALLES, J.; LOUREIRO, A. A. F. A comparison of Foursquare and Instagram to the study of city dynamics and urban social behaviour, **UrbComp'13, Proc. of the 2nd ACM SIGKDD International Workshop on Urban Computing**, Article No. 4, August 11-11, 2013, p. 1-8.

SILVA, T. H.; DE MELO, P. O. S. V.; ALMEIDA, J. M.; SALLES, J.; LOUREIRO, A. A. F. Uncovering properties in participatory sensor networks, **HotPlanet '12, Proc. of the 4th ACM international workshop on hot topics in planet-scale measurement**, 2012. p. 33-38.

ANEXO A QUESTIONÁRIO DA ENTREVISTA COM USUÁRIOS

Neste anexo, é reproduzido em formato de texto o que estava na página *web* do questionário. As etapas diferentes do questionário estão enfatizadas em negrito.

Descobrindo regiões semelhantes em Porto Alegre

Este questionário faz parte de um projeto de mestrado do INF-UFRGS e busca coletar opiniões sobre o aplicativo KANDOR, um mapa interativo de regiões da cidade de Porto Alegre. Através da visualização de regiões similares baseada em categorias de informação KANDOR exibe diferentes visões da dinâmica da cidade. Por exemplo, é possível ver as regiões mais populares, mais bem avaliadas ou mais frequentadas em determinados períodos do dia.

KANDOR apresenta cinco (5) categorias de informação que mostram regiões com locais com características semelhantes. Essas categorias são similaridade social, variedade de usuários, popularidade, avaliação e horários de pico.

O objetivo desta pesquisa é avaliar a relevância das categorias apresentadas e descobrir quais são as mais úteis para gerar visões da cidade.

Quando você escolhe uma categoria no KANDOR, o mapa interativamente exibe as regiões da cidade de acordo com a categoria selecionada. Ao clicar em um local específico, a região a que ele pertence e os outros locais semelhantes são destacados e uma tabela mostra características daquela região.

Você pode também selecionar na barra da esquerda um bairro da cidade para destacá-lo no mapa, facilitando a comparação entre as regiões encontradas pelo KANDOR e os bairros oficiais.

O tempo aproximado deste questionário é de 20 minutos e os resultados serão agrupados de forma anônima.

Informações Pessoais

Esta seção inclui algumas perguntas pessoais para entender melhor o perfil dos entrevistados. Os dados não serão divulgados.

1. Você mora em Porto Alegre?

- Sim
- Não

2. [Somente se respondeu NÃO pra pergunta acima] Já morou alguma vez em Porto Alegre? Por quanto tempo?

- Sim, por mais de 10 anos
- Sim, durante um período de 5 a 10 anos
- Sim, por menos de 5 anos
- Não

3. Qual sua idade?

menos de 18 anos

- 18-25 anos
- 26-30 anos
- 31-40 anos
- 41-60 anos
- acima de 60 anos

4. Qual sua profissão?

5. Qual seu nível de escolaridade?

- Ensino Fundamental Incompleto
- Ensino Fundamental
- Ensino Médio
- Ensino Superior
- Especialização
- Mestrado
- Doutorado
- Pós-Doutorado

6. Você possui um smartphone (celular inteligente com conexão com a internet)?

- Sim
- Não

7. Se respondeu sim pra pergunta acima, com qual frequência você utiliza redes sociais como Facebook, Instagram ou Foursquare?

- Diariamente
- 3 vezes por semana
- Uma vez por semana
- Menos de uma vez por semana
- Não uso

Familiarização com a Ferramenta

Para aprender a usar a ferramenta antes de responder as perguntas, esta seção apresenta um teste bem simples. Abra o link http://bit.ly/2iLhikw e você vai ser direcionado primeiramente para o modo teste. Atenção: não troque ainda para as outras categorias na barra da esquerda, somente o modo teste será usado aqui. No mapa, podem ser observadas duas regiões de Porto Alegre, cada uma com marcadores de uma respectiva cor somente para ajudar a diferenciar as regiões.

A seguir, observe os pontos destacados no mapa (passando o mouse em cima para ver o nome) e clique em pontos diferentes para ver as regiões encontradas. Ao examinar uma região, examine a tabela a direita do mapa para ver as características dessa região e comparar

as regiões existentes. Por exemplo, veja que informações são iguais e quais são diferentes. Com isto, responda as perguntas a seguir:

1. Quantas regiões foram encontradas no modo teste?

2. Quantos locais pertencem a região Verde?

3. Quantos locais pertencem a região Vermelha?

4. Analisando as informações da tabela para cada região, em quais informações elas se diferenciam?

Tutorial para uso do aplicativo KANDOR

O objetivo do aplicativo KANDOR é fornecer um mapa interativo com informações sobre a estrutura e a dinâmica das cidades. Para este experimento, foram obtidas publicações públicas e anonimizadas do Instagram na cidade de Porto Alegre com a localização de onde a foto foi tirada.

Foram gerados pontos no mapa representando locais que os usuários frequentaram, como restaurantes, universidades, parques, boates, etc. Ao passar o mouse sobre um ponto, é possível visualizar o nome do local.

Cinco (5) categorias de informações da cidade podem ser exploradas:
- Similaridade Social: lugares frequentemente visitados pelas mesmas pessoas
- Variedade de Usuários: quantidade de usuários diferentes
- Popularidade: quantidade de curtidas e comentários em posts de usuários
- Avaliação: notas em avaliações do Google
- Horários de Pico: quantidade de usuários por período do dia

Na barra de menu à esquerda, é possível trocar a visualização selecionando a categoria desejada. Na mesma barra, é possível também selecionar um bairro específico para marcá-lo no mapa e comparar com as regiões descobertas pelo KANDOR. Para navegar entre estas regiões, basta clicar em qualquer local no mapa (representado por um ponto colorido) e a região que ele está incluído e os locais semelhantes serão destacados. As cores dos pontos representam lugares similares, mas é possível que regiões diferentes tenham a mesma cor. Na dúvida, você pode clicar nos pontos para destacar as regiões em que eles se encontram.

A direita do mapa, você encontrará uma tabela com as características de cada região quando for selecionada. Ela exibe informações relativas a categoria selecionada e informam detalhes como proporção relativa de quantidade de curtidas, de número de usuários, de avaliações, etc.

O objetivo deste experimento é avaliar se as categorias propostas por KANDOR apresentam informações e regiões relevantes para entender a cidade.

A seguir, as próximas seções do questionário focam em bairros que você frequenta e a visualização deles. Cada sessão pergunta um tipo de bairro (moradia, trabalho/estudo e lazer) e pede que seja feita a análise das regiões descobertas próximas a ele.

Bairro onde você mora

1. Qual o bairro que você mora?

***** Passos Manuais *****

Acesse agora, por favor, o aplicativo KANDOR em http://bit.ly/2iLhikw. Lá, escolha na lista dos bairros de Porto Alegre na barra da esquerda o que você acabou de escrever na pergunta acima.

Para as próximas perguntas, você vai precisar clicar em cada uma das categorias (também na barra a esquerda) e analisar os resultados.

2. Para a categoria "Similaridade Social", avalie a seguinte frase: "As regiões encontradas representam lugares frequentemente visitados pelas mesmas pessoas."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

3. Para a categoria "Similaridade Social", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

4. Para a categoria "Variedade de Usuários", avalie a seguinte frase: "As regiões encontradas representam lugares com uma quantidade similar de usuários únicos."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

5. Para a categoria "Variedade de Usuários", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

6. Para a categoria "Popularidade", avalie a seguinte frase: "As regiões encontradas representam lugares com popularidade parecida (curtidas e comentários)."

7. Para a categoria "Popularidade", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

8. Para a categoria "Avaliação", avalie a seguinte frase: "As regiões encontradas representam localizações com qualidade semelhante."

O termo "qualidade" neste contexto é representado pela qualidade de serviço, produto ou atração oferecido pelo local.

9. Para a categoria "Avaliação", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

10. Para a categoria "Horários de Pico", avalie a seguinte frase: "As regiões encontradas representam lugares com atividade semelhante durante os mesmos períodos do dia."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

11. Para a categoria "Horários de Pico", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

12. Por favor, adicione quaisquer comentários que queira sobre o que achou sobre as categorias e os resultados encontrados.

Bairro onde você trabalha ou estuda

1. Qual o bairro que você trabalha ou estuda? (o que você passar mais tempo)

2. Qual frequência você visita o bairro acima?

	1	2	3	4	5	
Muito Pouco	○	○	○	○	○	Muito

***** Passos Manuais *****

Acesse agora por favor o aplicativo KANDOR em http://bit.ly/2iLhikw. Lá, escolha na lista dos bairros de Porto Alegre na barra da esquerda o que você acabou de escrever na pergunta acima.

Para as próximas perguntas, você vai precisar clicar em cada uma das categorias (também na barra a esquerda) e analisar os resultados.

3. Para a categoria "Similaridade Social", avalie a seguinte frase: "As regiões encontradas representam lugares frequentemente visitados pelas mesmas pessoas."

4. Para a categoria "Similaridade Social", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

5. Para a categoria "Variedade de Usuários", avalie a seguinte frase: "As regiões encontradas representam lugares com uma quantidade similar de usuários únicos."

6. Para a categoria "Variedade de Usuários", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

7. Para a categoria "Popularidade", avalie a seguinte frase: "As regiões encontradas representam lugares com popularidade parecida (curtidas e comentários)."

8. Para a categoria "Popularidade", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

9. Para a categoria "Avaliação", avalie a seguinte frase: "As regiões encontradas representam localizações com qualidade semelhante."

O termo "qualidade" neste contexto é representado pela qualidade de serviço, produto ou atração oferecido pelo local.

10. Para a categoria "Avaliação", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

11. Para a categoria "Horários de Pico", avalie a seguinte frase: "As regiões encontradas representam lugares com atividade semelhante durante os mesmos períodos do dia."

12. Para a categoria "Horários de Pico", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

13. Por favor, adicione quaisquer comentários que queira sobre o que achou sobre as categorias e os resultados encontrados.

Bairro que você frequenta por lazer

1. Qual o bairro que você mais frequenta por lazer? (restaurantes, bares, lojas, parques, etc.)

2. Qual frequência você visita o bairro acima?

	1	2	3	4	5	
Muito Pouco	○	○	○	○	○	Muito

***** Passos Manuais *****

Acesse agora por favor o aplicativo KANDOR em http://bit.ly/2iLhikw. Lá, escolha na lista dos bairros de Porto Alegre na barra da esquerda o que você acabou de escrever na pergunta acima.

Para as próximas perguntas, você vai precisar clicar em cada uma das categorias (também na barra a esquerda) e analisar os resultados.

3. Para a categoria "Similaridade Social", avalie a seguinte frase: "As regiões encontradas representam lugares frequentemente visitados pelas mesmas pessoas."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

4. Para a categoria "Similaridade Social", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

5. Para a categoria "Variedade de Usuários", avalie a seguinte frase: "As regiões encontradas representam lugares com uma quantidade similar de usuários únicos."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

6. Para a categoria "Variedade de Usuários", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

Concordo plenamente

7. Para a categoria "Popularidade", avalie a seguinte frase: "As regiões encontradas representam lugares com popularidade parecida (curtidas e comentários)."

8. Para a categoria "Popularidade", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

9. Para a categoria "Avaliação", avalie a seguinte frase: "As regiões encontradas representam localizações com qualidade semelhante."

O termo "qualidade" neste contexto é representado pela qualidade de serviço, produto ou atração oferecido pelo local.

10. Para a categoria "Avaliação", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

11. Para a categoria "Horários de Pico", avalie a seguinte frase: "As regiões encontradas representam lugares com atividade semelhante durante os mesmos períodos do dia."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

12. Para a categoria "Horários de Pico", avalie a seguinte frase: "As regiões que possuem pontos pra fora do bairro representam áreas que estão se propagando ou integrando com outros bairros."

	1	2	3	4	5	
Discordo plenamente	○	○	○	○	○	Concordo plenamente

13. Por favor, adicione comentários positivos/negativos sobre as categorias e os resultados encontrados.

Obrigado!

Suas respostas foram enviadas com sucesso. Muito obrigado pelo tempo e atenção para responder este questionário :)