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

BRENDA SALENAVE SANTANA

Extração e Aplicação de Indicadores no Processo de Recomendação de Recursos Urbanos Utilizando Dados Estruturados e Não-Estruturados

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

Orientador: Prof. Dr. Leandro Krug Wives

Porto Alegre
2019

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Santana, Brenda Salenave

Extração e Aplicação de Indicadores no Processo de Recomendação de Recursos Urbanos Utilizando Dados Estruturados e Não-Estruturados / Brenda Salenave Santana. – Porto Alegre: PPGC da UFRGS, 2019.

68 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2019. Orientador: Leandro Krug Wives.

1. Indicadores. 2. Recomendação. 3. Análise. 4. Dados. 5. Recursos Urbanos. 6. Informática Urbana. I. Krug Wives, Leandro. II. Título.

AGRADECIMENTOS

Nenhum estágio desta trajetória foi concluído sozinho, em todos os momentos pude contar com o apoio de muitas pessoas e sou grata a todas elas. Estes agradecimentos não poderão contemplar todos aqueles que de alguma forma contribuíram, mas para esses, fica a certeza que por mais que não expresso diretamente neste espaço, deixo meu carinho e mais sinceros agradecimentos.

Primeiramente, agradeço profundamente à minha mãe, Andria Salenave Santana, e ao meu pai, Carlos Roberto Soares Santana, pelo amor, pelo carinho e pela dedicação. Agradeço também de coração, a minha prima Bárbara Flores por todo incentivo e apoio dado, e aos demais familiares.

Gostaria de agradecer a todos meus amigos, em especial a Karina Lopes pelos incontáveis conselhos, ajudas e carinho ao longo de todos os anos de amizade, e também, por todos os momentos compartilhados. Agradeço ao meu amigo Lucas Ariel por todas as risadas, apoio e incentivos providos. Sou grata ainda por todo auxílio fornecido na execução deste trabalho, pelos meus amigos de Santa Maria, aos novos que conheci e aos antigos que reencontrei. Agradeço ainda àqueles que mesmo não estando mais aqui, sempre torceram e me motivaram a chegar a este ponto.

A todos os professores que durante este percurso proveram os ensinamentos necessários tanto para a realização deste trabalho quanto para a vida. Tenho a honra e a felicidade de dizer que tive grandes exemplos durante esta etapa percorrida. Agradeço em especial a professora Ana Trindade Winck por todas conversas, conselhos e ensinamentos. Faço também imenso agradecimento ao professor Leandro Krug Wives que me acolheu em seu grupo de trabalho, aconselhando e orientando o desenvolvimento deste trabalho, permitindo um enorme crescimento, sendo um grande exemplo em minha carreira acadêmica.

Obrigada ao Instituto de Informática da UFRGS. E um muito obrigada ao CNPq e à CAPES por financiarem parcialmente essa pesquisa.

E ainda, um agradecimento geral a todos aqueles que de alguma forma lutam e resistem pela nossa educação sabendo que a felicidade reside onde resiste a liberdade. Dessa forma, sigamos juntos pois somos muitos e não estamos sós.

RESUMO

Considerando o estudo do desenvolvimento de sistemas voltados a ambientes urbanos através da Informática Urbana, e tendo que dados referentes a tais de cenário encontram-se muitas vezes dispersos, em diferentes formas e estruturas e, em alguns casos, com procedência duvidosa, processos de recuperação e análise de informações tornam-se não-triviais. Nesse cenário, métodos capazes de extrair informações anteriormente desconhecidas ou não mensuradas e de valor para algum domínio são de fundamental importância. Diante de tal perspectiva, o principal objetivo desta pesquisa consiste em desenvolver uma abordagem capaz de extrair e analisar informações expressas em redes sociais baseadas em localização com o uso de Mineração de Textos, de modo a relacionar aspectos referentes a polaridade de informações e a confiabilidade dos perfis que as difundiram, bem como considerar o momento de avaliação, gerando indicadores a serem aplicados no processo de recomendação de recursos urbanos verificando tal influência ao estimar métricas de avaliação. Para tanto, procede-se a aplicação de uma metodologia baseada em premissas de análise de redes sociais, associada a aplicação de abordagens de *Web Mining* no processo de descoberta de conhecimentos e análise de dados. Como fonte de informações foi utilizado um conjunto de dados contendo 6600 observações coletadas no Foursquare, referentes à cidade de Gramado no Rio Grande do Sul, organizadas em 13 variáveis, além de informações complementares fornecidas pela plataforma DataViva. As características extraídas foram então aplicadas a algoritmos de recomendação baseados em vizinhança e em fatoração de matrizes, de modo a apurar métricas de acurácia com seu uso. Dos resultados obtidos, observa-se que, para algoritmos baseados em vizinhança, a abordagem proposta apresentou resultados melhores quando comparada à abordagem tradicional de avaliação. Entretanto, ao utilizar algoritmos baseados em fatoração de matrizes, as taxas de erro mantém-se com médias e desvios-padrão baixos. Os resultados obtidos foram comparados utilizando testes de Wilcoxon com 95% de confiança, o que permite concluir que esses retratam a não uniformidade na distribuição das amostras, evidenciando diferenças significativas entre os resultados obtidos.

Palavras-chave: Indicadores. Recomendação. Análise. Dados. Recursos Urbanos. Informática Urbana.

Extraction and Application of Indicators in the Urban Resource Recommendation Process Using Structured and Non-Structured Data

ABSTRACT

Considering the study of the development of systems directed to urban environments through Urban Computing, and that data related to such types of scenario are often dispersed in different forms and structures, some cases with doubtful origin, the recovery and information analysis processes become non-trivial. In this scenario, methods capable of extracting previously unknown information or not measured and of value for some domain are of fundamental importance. In this perspective, the primary objective of this research is to develop an approach capable of extracting and analyzing information expressed in social networks. The approach is based on location and uses of Text Mining, in order to relate aspects regarding the polarity of information and the reliability of the profiles that disseminated them, as well as considering the moment of evaluation, generating indicators to be applied in the process of recommending urban resources, verifying such influence when estimating evaluation metrics. For this purpose, a methodology based on social network analysis assumptions is applied, associated with Web Mining approaches in the process of knowledge discovery and data analysis. As a source, a dataset containing 6600 observations was collected at Foursquare, referring to the city of Gramado in Rio Grande do Sul. In this dataset, 13 variables were considered, and complementary information was provided by DataViva platform. The extracted features were applied to recommender approaches based on neighborhood and matrix factorization, and their use was measured in terms of accuracy. From the results, it is observed that the approach based on neighborhood algorithms presented better results when compared to the traditional evaluation approach. However, when using algorithms based on matrix-factorization, error rates are maintained with low standard means and standard deviations. The results obtained with the use of both metrics were compared using Wilcoxon tests with 95% confidence, which concludes that they portray the nonuniformity in the distribution of the samples, evidencing significant differences between the results obtained with the use of the approaches used.

Keywords: Indicators. Recommendation. Analytics. Data. Urban Resources. Urban Informatics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABC	Academia Brasileira de Ciências
DM	*Data Mining*
GPS	*Global Positioning System*
HTML	*HyperText Markup Language*
IE	*Information Extraction*
KDD	*Knowledge Discovery in Databases*
KDT	*Knowledge Discovery in Texts*
KNN	*K Nearest Neighbors*
LBSN	*Location-Based Social Networks*
NBR	Norma Brasileira aprovada pela Associação Brasileira de Normas Técnicas
OWL	*Web Ontology Language*
POI	*Point of Interest*
SR	Sistemas de Recomendação
SVD	*Singular Value Decomposition*
TF-IDF	*Term Frequency-Inverse Document Frequency*
WEB	*World Wide Web*
WM	*Web Mining*

SUMÁRIO

1 INTRODUÇÃO

Os dados gerados pelo intenso uso de diferentes tipos de sistemas computacionais multiplicam seu volume a cada dia. De acordo com estudos feitos pela Dell EMC (2017), existe uma estimativa que os dados mundiais cresçam a uma taxa de aproximadamente 40% ao ano durante a próxima década, de modo que o volume acumulado possa chegar a marca de 44 zettabytes ($44\text{x}2^{70}$ bytes) já no ano de 2020.

De acordo com Grus (2015), o cenário no qual vivemos atualmente está soterrado por dados, por exemplo, *websites* rastreiam todos os cliques dos usuários; os *smartphones* armazenam sua localização e velocidade a cada segundo; os pedômetros registram as batidas do coração, os movimentos, a dieta e até mesmo os padrões de sono; e a própria internet representa um enorme diagrama de conhecimento que contém uma enciclopédia de referências cruzadas (e.g., bases de dados sobre filmes, música, esportes e política). Assim, o aumento na quantidade de dados gerados por diferentes dispositivos acarreta em um volumoso conjunto de informações gerenciadas em um ambiente ubíquo[1].

Com a dispersão e o grande volume dos dados, a recuperação de informações relevantes deixa de ser trivial, tornando essencial a aplicação de técnicas de mineração de dados para a descoberta de conhecimento em documentos da *web*, conhecida como *Web Mining* (WM) de Conteúdo. Algumas técnicas exploradas para auxiliar no tratamento da sobrecarga de dados e extração de informações de textos têm sua origem na área de Descoberta de Conhecimento em Textos (do inglês *Knowledge Discovery in Texts* – KDT), que consiste na aplicação de técnicas de Descoberta de Conhecimento em Bases de Dados (do inglês *Knowledge Discovery in Databases* – KDD) sobre dados não estruturados (FELDMAN; DAGAN, 1995). Com a aplicação dessas técnicas torna-se possível, por exemplo, recomendar recursos urbanos aos cidadãos, que é um tema importante e atual.

Isso porque, com a expansão crescente da população (IBGE, 2019b), as cidades estão se tornando grandes centros urbanos e espera-se que elas sejam capazes de gerenciar seus diferentes tipos de recursos (e.g., atrações turísticas e gastronômicas). Conforme Thakuriah, Tilahun e Zellner (2017) explicam, a análise de elementos urbanos envolve o uso de uma gama de abordagens para entender os sistemas infraestrutural, físico e socioeconômico, e gerenciar setores complexos, como transporte, meio ambiente, saúde, habitação e economia. Desse modo, a denominada Informática Urbana é considerada uma área emergente e concentra-se na exploração e compreensão de sistemas urbanos,

[1] Neste trabalho, considera-se ubiquidade como o modo transparente que um sistema identifica possíveis fontes de dados (recursos) e conecta-se a elas, recomendando-as ou ignorando-as.

alavancando novas fontes de dados. São poucas as cidades brasileiras cujos recursos são conhecidos e projetados para seus habitantes, também conhecidas como “cidades inteligentes” (WEISS; BERNARDES; CONSONI, 2017).

A recomendação dos recursos disponíveis em ambientes ubíquos (e.g., cidades) requer a avaliação de diferentes amostras de informação (YING et al., 2012), não se limitando somente a dados estruturados, reforçando, assim, a pertinência do KDT. Por exemplo, a recomendação deve analisar um recurso urbano não somente pela informação em si, mas também sua origem (i.e., a confiabilidade da informação), sua polaridade (i.e., o quanto ela é positiva ou negativa) e a data da sua criação (i.e., se ela pode ser considerada ainda relevante dado o tempo decorrido de sua postagem).

Nesse contexto, este trabalho preocupa-se em estimular a utilização inteligente de recursos urbanos presentes em ambientes ubíquos, facilitando a integração dos cidadãos através do uso de métodos de recomendação mais acurados. Tais recursos, também denotados como equipamentos urbanos[2] ou pontos de interesse, podem então ser apresentados de modo adaptado a cada usuário, variando de acordo com o seu nível de conhecimento, necessidades e contexto.

Dessa forma, o objetivo geral deste trabalho consiste em desenvolver uma abordagem capaz extrair de analisar informações sobre recursos urbanos, expressas em redes sociais, baseadas em localização e com o uso de KDT, de modo a relacionar indicadores referentes a polaridade de informações e a confiabilidade dos perfis que as difundiram, bem como considerar o momento de avaliação, gerando indicadores [3] a serem aplicados no processo de recomendação de recursos urbanos verificando tal influência ao estimar métricas de avaliação. Para representação do conjunto estudado, foi utilizado um modelo ontológico desenvolvido sob o domínio de cidades visando integrar recursos para o desenvolvimento de aplicações de tecnologias de informação e comunicação, caracterizando um ambiente incipiente de cidades inteligentes, propiciando ainda a realização de inferências sob o cenário de estudos.

Os objetivos específicos deste trabalho são:

a) Investigar técnicas de KDT e de extração de informações, bem como sua aplicação

[2]De acordo com a ABNT NBR 9284, os equipamentos urbanos são “todos os bens públicos e privados, de utilidade pública, destinados à prestação de serviços necessários ao funcionamento da cidade, implantados mediante autorização do poder público, em espaços públicos e privados”. Ou seja, entende-se por equipamentos urbanos os elementos que compõem uma cidade, podendo abranger diferentes categorias, tais como: lazer, transportes, segurança, espaços comunitários e negócios locais.

[3]No contexto deste trabalho, o termo indicadores refere-se a métricas de avaliação geradas a partir do uso de características extraídas do conjunto de dados.

conjunta;

b) Analisar dados coletados pela mineração de conteúdo na Web (do inglês *Web Mining* – WM);

c) Aplicar tais dados a algoritmos de recomendação empregados em ambientes urbanos;

d) Extrair informações referentes aos recursos urbanos, considerando características (i.e., *features* de polaridade e confiabilidade) dos dados gerando indicadores a serem utilizados no processo de recomendação;

e) Verificar a acurácia da implementação por meio de métricas usadas por algoritmos de recomendação tradicionais;

f) Propor uma estimativa de avaliação para recursos urbanos, gerada a partir da agregação de diferentes indicadores extraídos dos conjuntos de dados;

g) Avaliar como o uso da estimativa gerada com o uso de indicadores (polaridade, confiabilidade e tempo) influencia no processo de recomendação de recursos de uma cidade através de métricas de acurácia (i.e., MAE e RMSE);

h) Utilização de um modelo de representação do conhecimento baseado em ontologias para o domínio de cidades, estruturando o conjunto de dados analisados permitindo a adaptação a diferentes cenários e ainda a realização de inferências sobre o conjunto.

Por se tratar de um tema recente e seus resultados dependerem da aplicação prática do KDT no contexto de cidades inteligentes, este trabalho pode ser classificado como uma pesquisa exploratória de natureza aplicada. Os objetivos do trabalho o caracterizam como uma pesquisa explicativa, uma vez que descreve a influência do KDT no processo de recomendação. Em relação aos procedimentos técnicos, foi adotada a pesquisa experimental para alcançar o propósito do trabalho, coletando dados de fontes como *Foursquare*[4] e plataformas de dados abertos como o *Data Viva*[5], seguindo as diretrizes indicadas por Fragoso, Amaral e Recuero (2011). Assim, foram desenvolvidas ferramentas de coleta e análise de dados com o enfoque em analisar o uso de indicadores de recomendação. A avaliação dos resultados foi feita de forma quantitativa, por meio de testes estatísticos de

[4]https://pt.foursquare.com
[5]http://www.dataviva.info/pt/

significância que permitem a comparação entre duas amostras pareadas (i.e., o processo de recomendação tradicional e o proposto por esta dissertação).

Para tanto, o trabalho está estruturado da seguinte forma: o Capítulo 2 descreve os fundamentos necessários para compreensão deste trabalho, bem como o estado da arte do tema abordado. O Capítulo 3 aborda os trabalhos que apresentam pesquisas análogas a esta proposta. O Capítulo 4 apresenta a proposta de desenvolvimento elaborada. O Capítulo 5 apresenta a aplicação das abordagens de análise do conjunto de informações para geração de indicadores auxiliares ao processo de recomendação de recursos urbanos. O Capítulo 6 detalha os experimentos aplicando resultados prévios das análises de dados em algoritmos recomendadores e os resultados alcançados diante dos cenários de testes utilizados para experimentações de validação da proposição exposta. Por fim, o Capítulo 7 apresenta a conclusão, as limitações e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo introduz os principais conceitos relacionados ao trabalho para facilitar o seu entendimento. No que diz respeito ao cenário de aplicação, a Seção 2.1 descreve a relevância das ditas Cidades Inteligentes. No que se refere a forma de representação de dados, a Seção 2.2 discorre brevemente sobre Modelagem Conceitual e Ontologias.

A Seção 2.3 conceitua a Descoberta de Conhecimento em Bases de Dados e possibilidades de aplicação deste. Já na Seção 2.4 aborda-se análise de dados, discorrendo sobre diferentes tipos e objetivos de pesquisa. Por fim, na Seção 2.5 são apresentados conceitos importantes de Sistemas de Recomendação (SR), bem como um breve desdobramento sobre Ontologias e suas aplicações dentro de tais sistemas.

2.1 Cidades Inteligentes

Conceitualmente, não há uma definição precisa e única do que é, de fato, uma cidade inteligente (do inglês, *smart cities*). Entretanto, a literatura afirma que para que uma cidade seja considerada inteligente, ela deve abranger três categorias principais de componentes conceituais: **(i)** tecnologia (i.e., infraestrutura tanto de *hardware* quanto de *software*), **(ii)** pessoas (e.g., criatividade, diversidade, educação), e **(iii)** instituição (e.g., governança, política) (NAM; PARDO, 2011). A partir de uma conexão entre estes três componentes, uma cidade é dita inteligente quando os investimentos do capital humano e social e a infraestrutura de tecnologia da informação incentivam o crescimento sustentável e melhoram a qualidade de vida da população a um custo acessível por meio de uma governança participativa (PÉREZ-MARTÍNEZ; MARTÍNEZ-BALLESTÉ; SOLANAS, 2013).

As cidades inteligentes foram recentemente registradas no livro da Acadêmia Brasileira de Ciências (ABC)[1] intitulado "Projeto de ciência para o Brasil: Cidades sustentáveis e inteligentes" como um tema de suma importância, apresentado na Seção Magna da ABC de (2017). Além da sua relevância no cenário atual, as cidades inteligentes surgem como um elo conceitual entre as áreas de ontologia e SRs. Desta forma, este trabalho foca em contribuir com a construção de uma cidade inteligente por meio do estudo de conceitos destas duas áreas para aperfeiçoamento do seu componente tecnológico *(i)*.

[1] http://www.abc.org.br/

2.2 Modelagem Conceitual e Ontologias

O conhecimento pode ser representado de diferentes maneiras, variando conforme a forma que um dado é descrito, representado e interpretado (ISOTANI; BITTENCOURT, 2015). Além da dificuldade de escolher um vocabulário que melhor identifique um conjunto de dados, um desafio da representação do conhecimento é retratar esses dados de maneira a aumentar a sua expressividade dentro do contexto em que foram criados e reduzir as possíveis ambiguidades que possam atrapalhar sua interpretação (ISOTANI; BITTENCOURT, 2015). Na tentativa de superar este desafio, a modelagem surge como uma forma de abstrair a realidade seguindo uma determinada conceituação.

A modelagem conceitual visa descrever a semântica, ou seja, os significados compreendidos, em um alto nível de abstração. (GUIZZARDI, 2005) afirma que a estrutura da conceituação do domínio deve ser acessível através de uma descrição explícita e formal de uma porção correspondente da realidade, em termos de um artefato concreto. Este artefato por vezes é denominado "Ontologia de Referência de Domínio", ou simplesmente, ontologia de domínio. Deste modo, a ideia é que esta ontologia deve ser construída com o único objetivo de fazer a melhor descrição possível do domínio na realidade, a um certo nível de granularidade e para um ponto de vista específico.

Ontologias podem ser utilizadas para capturar o conhecimento disponível no mundo e transformá-lo em um modelo processável automaticamente. Por conta disso, elas podem ser utilizadas na modelagem de informações do usuário e do contexto. De acordo com Isotani e Bittencourt (2015), com a informação e o conhecimento fragmentados em rede na era da *Web* de dados e com os recursos em constante evolução, desenvolver aplicações que façam uso de dados abertos não pode seguir o paradigma em que bases estáticas de dados são criadas para utilização em domínios restritos. Assim, vantagens que tais modelos de representação têm ao explorar o poder de representação e raciocínio das ontologias incluem:

a) Utilizar a expressividade da linguagem para descrever dados complexos;

b) Prover semântica formal aos dados, o que torna possível o compartilhamento e a integração de dados entre diferentes fontes;

c) Dispor de ferramentas de raciocínio tanto para checar a consistência das ontologias como para reconhecer que um determinado conjunto de instâncias de dados básicos e seus relacionamentos revelam a presença de uma caracterização de dados mais

abstrata (e.g., a atividade do usuário pode ser reconhecida automaticamente).

Segundo Gruber (1993), as ontologias são especificações formais e explícitas de uma conceituação compartilhada que pode servir como base de modelagem para diferentes fins de representação. Dentre suas aplicações, destaca-se o suporte oferecido à diferentes tipos de SR. Portanto, neste trabalho são utilizadas ontologias com o intuito de representar o conhecimento extraído a partir do conjunto observado, permitindo modelar diferentes cenários e, contribuindo assim para a realização de inferências sobre o conjunto e ainda possíveis extensões de aplicação.

2.3 Descoberta de Conhecimento

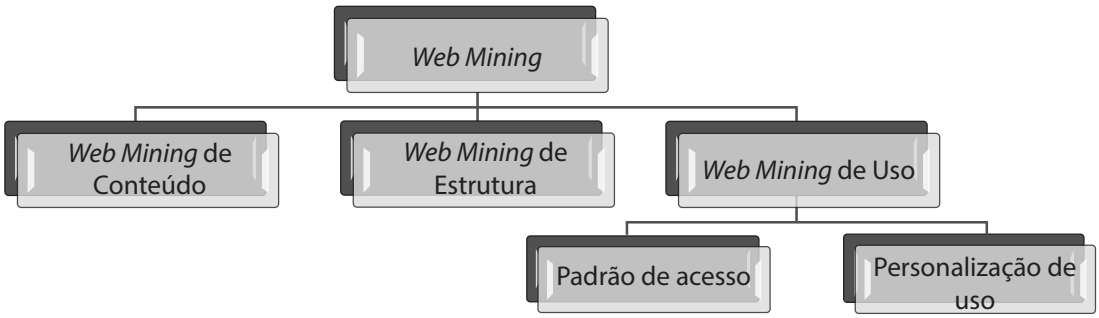
A análise dos volumes de dados torna-se cada vez mais complexa e inviável de ser realizada sem o suporte de ferramentas computacionais apropriadas. Assim, de forma a suprir tal necessidade, tem-se o estudo de Descoberta de Conhecimento em Bases de Dados. Em Fayyad, Piatetsky-Shapiro e Smyth (1996), a definição de KDD é dada como sendo um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados.

A etapa de Mineração de Dados (do inglês, *Data Mining* – DM) compreende a busca efetiva por conhecimentos úteis no contexto da aplicação de KDD. A mineração consiste no processo de explorar volumes de dados à procura de padrões coesos, de modo a encontrar relações sistemáticas entre variáveis, detectando assim, novos subconjuntos. Goldschmidt (2015), afirma que o valor atribuído ao dados armazenados está tipicamente ligado à capacidade de se extrair conhecimento de mais alto nível a partir deles, ou seja, informação útil que sirva para apoio à tomada de decisão e/ou para exploração e melhor entendimento do fenômeno que ocasionou. Os métodos de prospecção de informações possuem, portanto, o objetivo de transformar informações dispersas em um conjunto de dados em conhecimento.

A Web é uma das maiores e mais heterogêneas bases de dados disponíveis. De tal modo, mensurar o valor das informações contidas traz grandes desafios. Técnicas existentes de KDD destinam-se a auxiliar o processo de obtenção de conhecimento a partir de tais bases. Neste contexto, estas técnicas denominadas *Web Mining* (WM), agrupam ainda abordagens distintas, cada qual com seus métodos e ferramentas, sendo elas: *Web Mining* de Conteúdo, *Web Mining* de Estrutura e *Web Mining* de Uso, tal como ilustrado

na Figura 2.1.

Figura 2.1: Abordagens *Web Mining*.



Fonte: Goldschmidt (2015).

Para Chakrabarti (2002), *Web Mining* de Conteúdo constitui-se na aplicação de técnicas de Mineração de Dados para descoberta de conhecimento em documentos Web. Ou seja, o processo de mineração de conteúdos da Web consiste em analisar textos, imagens e outros componentes presentes em documentos Linguagem de Marcação de Documentos (do inglês, *HyperText Markup Language* – HTML), apresentando extensas informações, a fim de formatar visualmente o conteúdo. Para fins de aplicação desta técnica, tais informações são ignoradas e apenas seu conteúdo é considerado como base de dados para o processo de descoberta de conhecimento. Esta técnica é essencialmente utilizada como meio de facilitar o acesso ao conteúdo predominantemente desestruturado encontrado nestes tipos de documento. Dentre as principais utilizações, destacam-se a categorização automática de páginas HTML e indexação do conteúdo.

Já a mineração de estrutura de *links* da Web propõe-se ao estudo do relacionamento entre páginas da Web através de seus *hiperlinks*, ou seja, visa o desenvolvimento de técnicas para aproveitar o julgamento coletivo da qualidade de páginas Web que está implícito na estrutura de ligações. Alguns hiperlinks são utilizados de forma a organizar uma massiva quantidade de informações ou mesmo para facilitar a navegação do próprio site, já outros apontam para páginas de outros sites. Quando diversas referências apontam a uma mesma página, esta é considerada uma fonte de conteúdo de qualidade superior às demais páginas que não recebem tantos apontamentos. Atualmente, os principais motores de busca fazem uso desta informação para auxiliar o processo de ordenação de resultados de uma pesquisa.

Dentro das abordagens de WM, o *Web Mining* de Uso objetiva-se a realizar a análise de dados gerados por sistemas de informação baseados na Web relacionados com o

acesso de páginas. Tais análises envolvem a descoberta de padrões sobre dados muitas vezes armazenados em registros de servidores de aplicação. Assim, de acordo com Grace et al. (2011), o conteúdo de arquivos de *log* são comumente utilizados neste tipo de mineração. O processo de WM de Uso é dividido em três etapas principais: pré-processamento (tratamento dado ao conjunto de registros a serem analisados), mineração dos dados (aplicação dos algoritmos) e pós-processamento (análise e interpretação dos dados).

Por melhor se assemelhar aos objetivos deste trabalho, optou-se pela utilização híbrida das abordagens de *Web Mining* de Conteúdo e de Uso, ou seja, agregando a descoberta de conhecimentos em documentos dispostos na Web ao processo de realização da análise de dados gerados por tais sistemas de informação. A aplicação de conceitos referentes à WM de Conteúdo tem foco, então, na exploração dos diferentes tipos de informação encontrados na execução da proposta. Enquanto que os conceitos relativos à WM de Uso focam na análise de conteúdo gerada a partir do andamento do processo enfatizando a personalização de uso.

Tendo ainda que o desenvolvimento de plataformas de *hardware* e *software* para a Web, e também redes sociais permitiu a criação rápida de grandes repositórios de diferentes tipos de dados (AGGARWAL; ZHAI, 2012), tem-se diferentes técnicas exploradas para auxiliar no tratamento da sobrecarga de dados e extração de informações de textos. Este ramo tem sua origem na área de Descoberta de Conhecimento em Textos (do inglês *Knowledge Discovery in Texts* – KDT), que consiste na aplicação de técnicas de KDD sobre dados não estruturados (FELDMAN; DAGAN, 1995), sendo também tratado por estudos relacionados a mineração de textos. Segundo Aggarwal e Zhai (2012), embora dados estruturados geralmente sejam gerenciados com um sistema de banco de dados, dados textuais normalmente são gerenciados por meio de um mecanismo de pesquisa, devido à sua falta de estruturas definidas.

Assim, Hu e Liu (2012) expõe que técnicas de análise de texto podem ajudar a lidar eficientemente com dados textuais em mídias sociais para fins de pesquisa e negócios. Tendo que dados textuais em mídias sociais proporcionam *insights* sobre redes sociais e grupos que antes não eram possíveis em escala e extensão, esses autores expressam que a manipulação de tais dados apresentam diversos novos desafios devido às suas características distintas. Dentre as formas de análise de textos em mídias sociais se tem a análise de sentimentos.

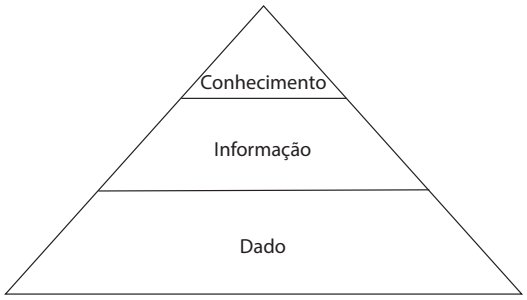
A análise de sentimentos é um processo de mineração do texto que identifica e extrai informações subjetivas no material de origem e auxiliando no entendimento senti-

mento social de uma marca, produto ou serviço. Liu e Zhang (2012) definem formalmente este conceito como o estudo computacional das opiniões, avaliações, atitudes e emoções das pessoas em relação a entidades, indivíduos, questões, eventos, tópicos e seus atributos. Deste modo, de acordo com os autores se tem que devido ao crescimento explosivo das mídias sociais (e.g, revisões, discussões em fóruns, blogs e redes sociais) na Web, indivíduos e organizações utilizam cada vez mais opiniões públicas expressas nesses meios para a tomada de decisões, sendo o processo de obtenção de informações relevantes nada trivial. Tem-se desta forma que em alguns casos torna-se difícil identificar sites relevantes e resumir com precisão as informações e opiniões contidas neles. Além disso, Liu e Zhang (2012) ressaltam ainda que a análise humana da informação textual está sujeita a vieses consideráveis, de forma que muitas vezes o foco de atenção concentra-se em a opiniões que são consistentes com as suas próprias preferências.

2.4 Análise de Dados

Em Goldschmidt (2015) o autor apresenta a hierarquia existente entre Dado, Informação e Conhecimento, tal como ilustrado na Figura 2.2. Esta busca representar o valor agregado aos dados conforme o (pré-)processamento que é realizado, de modo que quanto mais alto o nível alcançado na pirâmide maior é o valor que tal dado contém.

Figura 2.2: Hierarquia entre Dado, Informação e Conhecimento.



Fonte: Goldschmidt (2015).

Dados podem ser classificados entre os estruturados e os não estruturados. Dados estruturados são aqueles que contém uma organização para serem recuperados. São aqueles armazenados dentro de uma estrutura que permite seu entendimento de modo simplificado, pois estão armazenados de forma sistemática. Tal organização é geralmente feita por colunas e linhas, podendo variar de acordo com a fonte de informações utilizada.

Diferentemente deste, dados não estruturados não apresentam uma organização

explícita em disposição. Este tipo refere-se a dados que não podem ser organizados em linhas e colunas, como por exemplo vídeos e textos. Geralmente estes são de difícil acesso e recuperação, e muitas vezes não dispõem de componentes necessários para identificação de tipo de processamento e interpretação, tornando o seu uso um desafio.

O processo de análise de dados passa por diferentes técnicas que se definem pelos níveis de aprofundamento e objetivos finais. De acordo com Davenport (2013), existem três principais tipos de análise:

a) Análise Preditiva: A partir da identificação de padrões busca identificar tendências futuras, tal como é realizado em SRs. De acordo com Shmueli e Koppius (2011) este tipo de análise auxilia na criação de modelos, também desempenhando um papel importante ao lado da modelagem explicativa na construção e no teste de teorias.

b) Análise Prescritiva: Traça possíveis consequências de cada ação. Busca definir qual escolha mais efetiva em determinada situação. Segundo Davenport (2013), este tipo de análise utiliza modelos para determinar melhores comportamentos e ações.

c) Análise Descritiva: De acordo com Evans e Lindner (2012), esta análise faz o uso de dados para entender o desempenho de ações passadas e atuais e assim tomar decisões informadas. Segundo este autor, a análise descritiva é o tipo mais comumente utilizado e compreendido. Tais técnicas categorizam, caracterizam, consolidam e classificam dados para convertê-los em informações úteis para fins de compreender e analisar o desempenho de negócios.

2.5 Sistemas de Recomendação

Sistemas de recomendação (SR) são definidos de forma abrangente por Ricci, Rokach e Shapira (2011) como modelos, técnicas e ferramentas de software usadas para prover sugestões de itens a serem utilizados por um usuário. Dada tal definição tem-se que pelo menos duas dimensões de informação são utilizadas em SR tradicionais, sendo elas, os perfis do usuário e do item a ser recomendado. Com o poder de expressão de conhecimento de domínio, em Gemmis et al. (2015) evidencia-se que diversas abordagens têm sido propostas para incorporar conhecimento ontológico em SR. Em Jannach et al. (2010) afirma-se que sistemas de recomendação podem ser classificados em: baseados em conteúdo, baseados em filtragem colaborativa e abordagem híbrida.

Abordagens baseadas em conteúdo, de acordo com Lops, Gemmis e Semeraro (2011), buscam recomendar itens semelhantes àqueles aos quais determinado usuário aprovou anteriormente. Ou seja, o processo básico executado por um recomendador baseado em conteúdo consiste em combinar os atributos de um perfil de usuário no qual preferências e interesses são armazenados, com os atributos de um objeto de conteúdo (item), para recomendar ao usuário novos itens de interesse. De acordo com Tuzhilin e Adomavicius (2005) esta abordagem tem sua origem em pesquisas de recuperação de informações e de filtragem de informações. Assim, segundo os autores, dados os avanços significativos realizados por estas áreas, muitos dos sistemas baseados em conteúdo se concentram em recomendar itens que contêm informações textuais. Entretanto de acordo com estes, a abordagem apresenta uma série de limitações de uso.

Em Isinkaye, Folajimi e Ojokoh (2015) tem-se que a técnica de filtragem colaborativa é a mais madura e a mais comumente implementada. Esta, de acordo com os autores, realiza a recomendação de itens identificando outros usuários com preferências semelhantes, utilizando suas propensões para recomendar itens a usuários ativos. Em concordância com Ricci, Rokach e Shapira (2015), tem-se que a similaridade na preferência de dois usuários é calculada com base na semelhança do histórico de classificação dos usuários. Por tal razão, é possível encontrar na literatura referências à filtragem colaborativa como "correlação de pessoas para pessoas".

Conforme Goldschmidt (2015), sistemas de recomendação colaborativa podem ser organizados de acordo com a seguinte classificação: recomendação baseada em usuário e baseada em item. De acordo com Sarwar et al. (2001), nesta primeira técnica analisa-se primeiramente a matriz de item de usuário para identificar relações entre diferentes itens e, em seguida, utiliza esses relacionamentos para calcular indiretamente recomendações para usuários. Já as abordagens baseadas no usuário, buscam analisar semelhanças entre estes e as utiliza para calcular indiretamente recomendações de itens análogos.

Em Tuzhilin e Adomavicius (2005) ressalta-se que tanto abordagens baseadas em conteúdo quanto em filtragem colaborativa utilizam a similaridade de cosseno, tendo fundamentos na literatura referente à recuperação de informações. No entanto, em sistemas de recomendação baseados em conteúdo, ele é usado para medir a similaridade entre vetores de pesos da frequência do termo-inverso da frequência nos documentos (do inglês *term frequency–inverse document frequency* – TF-IDF), enquanto que SR baseados em filtragem colaborativa, estimam a semelhança entre os vetores de classificações reais especificadas pelo usuário.

Tem-se ainda que a abordagem híbrida se utiliza de uma combinação das estratégias anteriores para solucionar problemas comuns dessas, tais como o *coldstart*, onde o sistema tem dificuldades de prover boas recomendações para novos usuários e novos itens, bem como o *overfitting*, onde o usuário acaba recebendo conteúdo reincidente, não permitindo diversificação de interesses.

De acordo com Amatriain e Pujol (2015) sistemas de recomendação usualmente aplicam técnicas e metodologias de outras áreas, tais como Interação Humano Computador Recuperação de Informação. Assim, segundo Amatriain (2013), a maioria desses sistemas tem em seu núcleo um algoritmo que pode ser entendido como uma instância específica de um processo de Mineração de Dados.

Neste trabalho, a aplicação de sistemas de recomendação se dá com a utilização de dados provenientes de redes sociais baseadas em localização. Em Bao et al. (2015), os autores exploram o fato de que recentes avanços nas técnicas de localização aprimoraram fundamentalmente os serviços de redes sociais, permitindo que os usuários compartilhem seus locais e conteúdos relacionados à localização, como fotos e anotações geo referenciadas em LBSNs.

Zheng e Zhou (2011) definem que esse tipo de rede social não significa apenas adicionar uma localização a uma existente para que as pessoas na estrutura social possam compartilhar informações incorporadas à localização, mas também consiste na nova estrutura social composta de indivíduos conectados pela interdependência derivada de suas localizações no mundo físico, bem como seu conteúdo de mídia marcado por localização, como fotos, vídeo e texto. Para os autores, a localização física consiste na localização instantânea de um indivíduo em um determinado registro de data e hora e no histórico de localização que um indivíduo acumulou em um determinado período. Além disso, a interdependência inclui não apenas que dois usuários co-ocorram no mesmo local físico ou compartilham histórias de localização semelhantes, mas também o conhecimento, (e.g., interesses comuns, comportamentos e atividades, inferidos da localização de um indivíduo e do histórico e dados).

2.6 Considerações sobre o Capítulo

O estudo e desenvolvimento de propostas que se destinam a análise de informações provenientes da Web para extração de características de interesse do usuário, vem sendo amplamente trabalhado na área de recuperação de informações. O desenvolvimento de

pesquisas na área se deve ao fato do crescimento do volume de dados trafegados e do valor das informações ali contidas. É possível encontrar de modo amplo na literatura, diversos métodos e algoritmos propostos que podem ser utilizados para o aprimoramento em métricas de acurácia de recomendação.

Ao considerar o volume de dados gerados em um ambiente de cidade inteligente, surge a necessidade de representação. Logo, o uso de ontologias auxilia a capturar o conhecimento disponível no ambiente, convertendo-o em um modelo interpretável e processável. Assim, a utilização de abordagens de descobertas de conhecimento propiciam técnicas para extração e mineração de diferentes tipos de dados. Para tanto, conjuntamente a tais abordagens neste trabalho aplicam-se métodos de análise de dados, buscando explorar com maior completude o conjunto de dados observados. Tem-se então, a partir do uso de tais meios, uma forma de prover dados do cenário estudado em maior completude ao modelo ontológico utilizado como base.

Dessa forma, realizando uma análise sobre os dados extraídos torna-se possível extrair *features* de interesse, de modo que essas possam auxiliar na recomendação de recursos dispostos no ambiente. Conjunto ao uso dessas informações, considera-se então empregar uma métrica de avaliação de pontos de interesse gerada a partir do resultado de análises considerados como indicadores em tal processo de recomendação, para aprimorar a acurácia desses sistemas.

3 TRABALHOS RELACIONADOS

Uma vez que o volume de informações cresce de maneira constante, torna-se comum encontrar dados em diversos formatos e estruturas, dispersos e em alguns casos, de origem inconsistente. Com isso, o processo de sua recuperação torna-se não trivial. Neste Capítulo são descritos trabalhos correlatos os quais realizaram pesquisas análogas aos objetivos do presente trabalho. Assim, são apresentados trabalhos resultantes de pesquisas sobre a descoberta de conhecimentos em redes sociais baseadas em localização, bem como a recomendação de itens urbanos a partir do uso de tais redes.

Em Goldschmidt (2015) o autor expõe uma ligação existente entre *Big Data* e KDD. Para tanto, é retratado que ambos os termos podem ser vistos sob três perspectivas, sendo estas:

a) Perspectiva Científica: a qual abrange pesquisas voltadas à novos recursos para o processamento dos dados (coleta, preparação, integração, armazenamento, recuperação e análise), bem como ao aprimoramento de recursos já existentes.

b) Perspectiva Tecnológica: a qual refere-se ao conjunto e a aplicação de recursos disponíveis a diferentes contextos.

c) Perspectiva Mercadológica: a qual compreende iniciativas voltadas a exploração de resultados obtidos nos cenários de aplicação.

Assim, dada a interseccionalidade entre as áreas, tem-se que estas podem ser utilizadas de modo conjunto em diferentes contextos. Como por exemplo, em Thakuriah, Tilahun e Zellner (2017), os autores descrevem fontes de *Big Data*, consideradas por estes como emergentes, e seu uso na pesquisa urbana também sob a perspectiva de desafios de aplicação. De acordo com Thakuriah, Tilahun e Zellner (2017), o termo *Big Data*, é utilizado para descrever um amplo espectro de dados observacionais gerados por meio de atividades de diferentes tipos (i.e., transacionais, operacionais, de planejamento e sociais), que não necessariamente foram inicialmente projetadas para pesquisa. Assim, este relata que o *Big Data* no contexto urbano se tornou estreitamente associado a dados de sensores (i.e., internet das coisas) ou gerados socialmente (i.e., mídia social). Deste modo, dada a estrutura e condições de acesso associadas a tais dados, seu uso para pesquisa e análise torna-se significativamente complicado.

Segundo Mehmood et al. (2016), no ciclo de vida *Big Data* a etapa de geração de dados existente, reforça que as informações podem ser providas por fontes diversas e

distribuídas. Em adição a isto, em Strohbach et al. (2015) relata que aumento da urbanização e a implantação de comunicações entre máquinas trazem aumentos significativos nos dados gerados por aplicações IoT implantados no cenário de cidades inteligentes.

Ao se tratar de cenários urbanos, tem-se então que o volume de informações pode ser significativo, possuindo seu acesso dificultado dada a possível dispersão de dados. Em redes sociais baseadas em localização (do inglês, *Location-Based Social Networks* – LBSNs), usuários podem realizar *check-in* e deixar dicas comentando sobre um local, deste modo, LBSNs podem reunir de uma melhor forma informações referentes a pontos de interesse. De modo que tais fontes de dados se mostram heterogêneas, descrevem ainda a preferência dos usuários pelos locais. Segundo Yang et al. (2013), em diversas pesquisas considera-se apenas o *check-in* para recomendação de pontos de interesse (do inglês, *Points od Interest* – POI), de modo que *reviews* de usuários são pouco estudados. Assim, em seu trabalho os autores buscam aprimorar a recomendação de tais elementos modelando a preferência de localização do usuário, bem como o algoritmo de recomendação. Ao analisar dados provenientes de LBSNs, Yang et al. (2013) utiliza um método de análise sentimentos não-supervisionado baseado em dicionário para o processamento de dicas fornecidas.

De tal forma, o autor propôs um modelo híbrido de recomendação, onde buscou-se combinar *features* como a preferência extraída de *check-ins* e dicas textuais que são processadas usando técnicas de análise de sentimentos, com o pressuposto de que *reviews* fornecidos sobre determinado local podem auxiliar em uma melhor caracterização de preferências. Com base no número de confirmações de localização e nas pontuações de sentimento extraídas, foi elaborada uma matriz de preferência do local do usuário. Yang et al. (2013) propôs então um algoritmo de fatoração de matrizes, considerando a influência social do usuário e a influência da similaridade de local na recomendação. Ao realizar a avaliação de tal abordagem, foram atingidos bons resultados na acurácia (MAE e RMSE), demonstrando que o modelo de preferência híbrida pode caracterizar melhor a preferência dos usuários e verificar a eficácia e a eficiência da abordagem proposta de fatoração de matriz social baseada em localização.

Em Zhang (2016), o autor propôs um sistema de recomendação centrado em grupo, no domínio de Sistemas Ciber-Físico-Sociais (CPSS), que consistiu na descoberta de grupos orientados por atividades. O sistema proposto centrado em grupo, chamado GroRec, integra tecnologias sociais, móveis e de *big data* para fornecer serviços de recomendação eficazes em CPSSs. A abordagem centrada em grupo baseada na similaridade

comportamental entre os usuários busca diminuir a complexidade de sistemas convencionais de recomendação centradas no indivíduo.

Assim, em seu trabalho Zhang et al. (2016), utiliza um método de quantificação de compensação emocional baseado na análise de sentimento, proposto para revisar as avaliações do usuário para melhorar a objetividade dos dados de classificação. A abordagem aplicada é baseada em três aspectos, ou seja, classificações, interesses e relações sociais, sendo proposta para extrair de forma abrangente as preferências do grupo. Experimentos realizados, demonstram que o sistema de recomendação proposto por estes se mostrou eficiente, objetivo e preciso, fornecendo assim uma base sólida para a computação personalizada no paradigma do CPSS. De acordo com os resultados apresentados, o desempenho do sistema se mostra superior ao das abordagens convencionais baseadas em filtragem colaborativa e fatoração de matrizes.

Em Ying et al. (2012) tem-se que mesmo com um aumento no número de pesquisas sobre recomendação de pontos de interesse urbanos, e com a série de técnicas de recomendação baseadas na sociedade propostas na literatura, grande parte destas aplicações baseia-se apenas no comportamento de *check-in* do indivíduo ou de pessoas de seu convívio. Assim, o autor afirma que isso leva a uma lista de recomendação geralmente restringida na área de convivência dos usuários ou conhecidos deste. De modo a lidar com informações contextuais e informações do ambiente, o autor propôs uma abordagem denominada Urban POI-Mine (UPOI-Mine) integrando LBSNs para recomendação de POIs urbanos com base nas preferências do usuário e propriedades de localização, simultaneamente.

Assim, Ying et al. (2012) sugere a UPOI-Mine como forma de elaboração de um preditor baseado em árvore de regressão no espaço de *check-in* normalizado, de modo a suportar a predição de interesse do POI relacionado à preferência de cada usuário. Com base em dados de LBSN, foram extraídas informações referentes ao fator social (*check-ins* entre amigos similares ao usuário no ponto de interesse), preferência individual (predisposição do usuário para com marcações semânticas do sistema) e popularidade do POI para construção de modelo. Por meio de avaliações experimentais em um conjunto de dados reais de rede social (Gowalla), onde a abordagem proposta apresentou bons resultados em suas métricas de acurácia.

Em Waga, Tabarcea e Franti (2013), apresenta-se um sistema de recomendação contextual personalizada, o qual buscou estimar dados relevantes baseados em localização da coleção de usuários para recomendação. O sistema utilizado fornece a recomendação

de três tipos de itens, sendo estes de serviços, fotos e rotas de GPS (*Global Positioning System*) que são considerados pontos de interesse no ambiente do usuário. O algoritmo apresentado utiliza três parâmetros como entrada, sendo o primeiro utilizado para identificar o usuário para o qual a recomendação é personalizada, o segundo parâmetro é a localização deste, e o terceiro parâmetro é o horário da solicitação de recomendação.

Dessa forma, em Waga, Tabarcea e Franti (2013) os itens são pontuados usando diversos critérios baseados em aspectos de relevância (priorizando dados recentes). O processo de seleção e pontuação é executado para serviços, fotos e rotas separadamente. Os resultados alcançados pelo autor demonstram que o método é capaz de proporcionar a seleção de itens relevantes para recomendação, destacando que a influência temporal possui grande importância visto que determinadas rotas e lugares são mais visitadas em diferentes épocas do ano.

Segundo Gao et al. (2015), embora o trabalho existente de recomendação de POI em redes sociais baseadas em localização seja capaz de descobrir padrões espaciais, temporais e sociais do comportamento de *check-in* do usuário, o uso de informações de conteúdo (como por exemplo propriedades do POI, interesses do usuário e indicações de sentimento) ainda não foi sistematicamente estudado. Sendo que a recomendação de POI é uma tarefa que facilita a exploração urbana dos usuários e os ajuda a filtrar os POIs desinteressantes para a tomada de decisões, em seu trabalho os autores modelam os três tipos de informação sob uma estrutura unificada de indicação de POI com a consideração de sua relação com as ações de *check-in*.

Os resultados experimentais alcançados sobre estudos realizados um conjunto de dados da rede Foursquare no trabalho de Gao et al. (2015), exibem a importância das informações de conteúdo para explicar o comportamento do usuário e demonstram seu poder de melhorar o desempenho das indicações de POI em LBSNs. Deste modo, a modelagem dos fatores de propriedades do POI, interesses do usuário e indicações de sentimento, utilizada pelos autores, representam indicadores relevantes para o processo de recomendação de recursos urbanos fundamentando-se em dados de redes sociais baseadas em localização.

Bellini et al. (2014) propõe um sistema de recomendação utilizando a conciliação de dados públicos e privados de aspectos relacionados a cidades inteligentes. De acordo com os autores, o sistema permite gerenciar grandes volumes de dados provenientes de uma variedade de fontes, considerando tanto dados estáticos quanto dinâmicos. Realizando o mapeamento dos dados para um Modelo de Conhecimento para Cidade

(KM4City), e armazenando-os em um RDF-Store, onde estão disponíveis via consultas SPARQL para fornecer novos serviços aos usuários por meio de aplicações direcionadas a administração pública e empreendimentos, Bellini et al. (2014) apresenta o processo adotado para produzir a ontologia nesse contexto. Os autores dessa pesquisa apresentam, ainda, mecanismos adotados para a verificação, reconciliação e validação de dados, contribuindo assim para o desenvolvimento de ontologias fidedignas ao cenário modelado.

Em Komninos et al. (2016), argumenta-se que o impacto de aplicações desenvolvidas no contexto de cidades inteligentes decorre essencialmente da ontologia utilizada e, suplementarmente, de recursos inteligentes de tecnologia e programação. Por consequência, autores defendem que o uso de ontologias possui implicações significativas para o design e o desenvolvimento de aplicações para o cenário. Assim, o aperfeiçoamento de sua eficácia de uso e melhoria do impacto decorre da prioridade dada ao projeto de sua ontologia e da relação com a ontologia geral da cidade inteligente e as classes e propriedades contidas na aplicação. Os autores afirmam ainda que uma série de estratégias podem vir a contribuir para aplicações mais bem-sucedidas e de alto impacto, bem como podem priorizar atividades que afetam o sistema de inovação da cidade.

3.1 Resumo do Capítulo

Dentre os trabalhos correlatos aqui apresentados, tem-se uso de diferentes técnicas e ferramentas que buscam o processamento e análise de conteúdos que remetam ao melhoramento de recomendações a partir do uso de dados de redes sociais baseadas em localização. Dessa forma, o uso de *features* extraídas a partir de métodos de análise de dados, aplicado a sistemas de recomendação, tem se mostrado eficaz quanto ao aprimoramento dos resultados avaliados em diferentes modelos.

Entretanto, dentre os trabalhos analisados que utilizaram características extraídas dos conjuntos de dados, destaca-se em todos os casos a utilização de elementos auxiliares no processo de recomendação, limitou-se a apenas um aspecto. Diferente de pesquisas correlatas anteriores, ao entender que diferentes atributos podem possuir maior ou menor relevância para o usuário, neste trabalho buscou-se abordar a combinação de múltiplas *features* extraídas a partir de uma rede social baseada em localização. E, dessa forma, realizar a aplicação da estimativa gerada a algoritmos de recomendação, visando atingir melhores taxas de acurácia na recomendação de recursos urbanos.

Tendo ainda que o uso de ontologias se mostra de grande importância para o de-

senvolvimento de aplicações voltadas ao cenário de cidades inteligentes, assim como nos trabalhos correlatos, este tem como base o uso de um modelo de representação do conhecimento assim expresso. Tendo em vista que este cenário passa por constantes mudanças, uma vez que ambientes ubíquos passam por alterações a todo instante, neste trabalho apresenta-se também a população automatizada da ontologia na qual o desenvolvimento foi baseado. Deste modo, a atualização do modelo torna-se capaz de manter-se facilmente atualizada, servindo como possível aspecto de agregação aos algoritmos de recomendação utilizando ainda as *features* extraídas.

4 ABORDAGEM PROPOSTA

Este Capítulo apresenta a proposta de extração de *features* relevantes através de abordagens de análise de dados para aplicação a sistemas de recomendação. Assim, descreve-se o desenvolvimento do presente trabalho e da proposta de extração de indicadores de recomendação de recursos urbanos com utilização de análise de dados. Logo, este Capítulo encontra-se organizado da seguinte forma: a Seção 4.1 expõe a metodologia utilizada como base para condução dos estudos realizados; a Seção 4.2 apresenta a estrutura sobre a qual se baseia o desenvolvimento deste trabalho, assim, descreve como ocorrem os processos principais envolvidos, sendo esses referentes a coleta, análise dos dados, recomendação gerada a partir desses e a avaliação de tais métodos; por fim, a Seção 4.3 destina-se a descrição dos algoritmos de recomendação utilizados.

4.1 Metodologia

Este trabalho segue a abordagem descrita em Fragoso, Amaral e Recuero (2011) para o estudo de redes sociais. Com o uso desse processo, as autoras traçam pontos principais de condução de métodos de pesquisa envolvendo redes sociais na internet. Baseando-se em premissas da 'Análise de Redes Sociais', são determinadas duas etapas principais, sendo essas: a delimitação do objeto e dados.

Dentro da delimitação do objeto, busca-se traçar a determinação de uma rede social a partir do objeto do pesquisador, ou seja, dos elementos os quais busca-se analisar a fundo. Desse modo, em Fragoso, Amaral e Recuero (2011) define-se que é necessário selecionar o objeto e a forma de coleta antes de iniciar a análise. Assim, neste trabalho escolheu-se como objeto de estudos o Foursquare, dada sua política de uso de dados e permissão de uso para fins acadêmicos, sendo ainda representativo como fonte de dados de redes sociais baseadas em localização.

Ao se tratar dos dados observados, foi então desenvolvido um *crawler* para coleta de informações dispostas na plataforma, onde se tem que os dados são estruturados de forma dinâmica, uma vez que apresenta comportamentos como cooperação, pequenos *clusters*, etc. A relação entre os usuários é tida como sendo uma interação de manutenção, de modo que essas visam apenas contribuir fornecendo alguma informação e não com o objetivo de aumentar a intimidade entre os usuários (FRAGOSO; AMARAL; RECUERO, 2011).

4.2 Estrutura

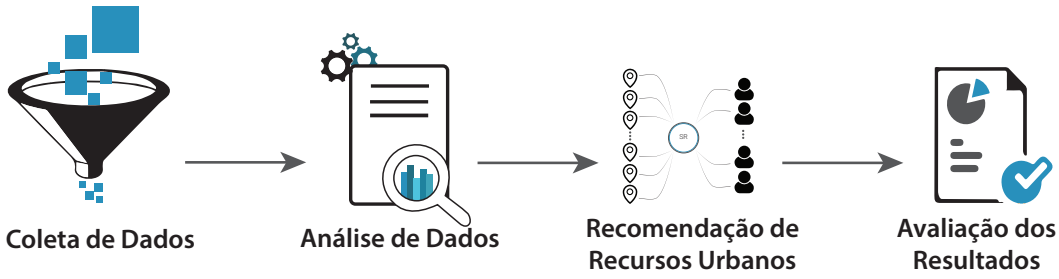
Dada a quantidade de recursos dispostos em uma cidade, o volume de informações gerado é extenso. Assim, a escolha de recursos por parte do cidadão torna-se não trivial devido a quantidade de dados e informações disponíveis. Plataformas voltadas ao fornecimento de informações e opiniões de conteúdos relacionados ao turismo, tais como redes sociais baseadas em localização, oferecem mecanismos de filtragem, inclusive com avaliações (*ratings*) e comentários (*reviews*).

Entretanto, ao realizar uma consulta, esses muitas vezes não consideram o contexto do usuário e seus interesses, nem mesmo realizam recomendações combinando tais fatores. Ou em algumas situações, como é o caso da rede social analisada, são considerados aspectos incrementais na avaliação, porém não são apresentadas explicações de recomendação aos usuários. Dessa forma, o presente trabalho parte do pressuposto de que agregar informações extraídas de elementos textuais (*reviews*) com dados estruturados (propriedades dos recursos, informações pessoais, de interesse e de contexto dos usuários) auxiliaria nesse processo de escolha.

A estruturação da proposta apresenta-se dividida em quatro módulos principais: *Coleta de Dados*, *Análise dos Dados*, *Recomendação de Recursos Urbanos* e *Avaliação*, tal como ilustrado na Figura 4.1. Neste capítulo faz-se a descrição geral da proposta. Detalhes de como ela foi efetivamente implementada e utilizada encontram-se no próximo capítulo.

Com o uso de tal infraestrutura, tem-se então o objetivo de avaliar o uso de indicadores auxiliares no processo recomendação, bem como a análise de aspectos relacionados a dados dispostos em diferentes formas e estruturas.

Figura 4.1: Estrutura



Fonte: Acervo pessoal.

4.2.1 Coleta de Dados

De acordo com Norvig e Russell (2014), o objetivo da extração de informações (do inglês *Information Extracion* – IE) é de processar a linguagem natural expressa textualmente para recuperar ocorrências de uma determinada classe de objetos ou eventos e ocorrências de relações entre eles. Assim, a IE é o processo de extrair informações específicas (pré-especificadas) de fontes textuais. Desse modo, agregou-se os objetivos desta a *Web Mining* visando realizar o processo de extrair e analisar informações extraídas de fontes *Web*.

Neste trabalho, focou-se no estudo de dados disponíveis na rede social baseada em localização Foursquare. Para tanto, foi desenvolvido um *crawler* a fim de coletar informações de redes colaborativas de avaliação de recursos urbanos. Focado principalmente em *Web Mining* de Conteúdo, o coletor desenvolvido em Python foi implementado de maneira *ad-hoc* visando capturar informações tais como: nome do estabelecimento, tipo de estabelecimento, localização (latitude e longitude), valor médio das avaliações desse, número total de avaliações, comentários opinativos sobre o estabelecimento, número de votos favoráveis ou contrários aos comentários, e ainda o número de avaliações feitas no site anteriormente pelo usuário.

De modo a melhor estruturar a representação dos dados obtidos, elaborou-se um modelo representação do conhecimento, baseado em ontologia no domínio de cidades onde se pudesse reunir tais informações. Dessa forma foi desenvolvida uma ontologia para cidades para tal fim, a qual também é descrita em Santana, Oliveira e Wives (2018). Foram ainda extraídos dados do portal DataViva, o qual é uma plataforma aberta e gratuita de visualização de dados sociais e econômicos do Brasil, da referida cidade de interesse em vista de dar maior completude ao modelo.

Em Wimalasuriya e Dou (2010), apresenta-se que a extração de informação baseada em ontologias emergiu como subcampo da IE, onde ontologias são utilizadas pelo processo de extração de informações e a saída é geralmente apresentada por meio de uma ontologia. A seguir apresenta-se a forma como o modelo desenvolvido encontra-se organizado.

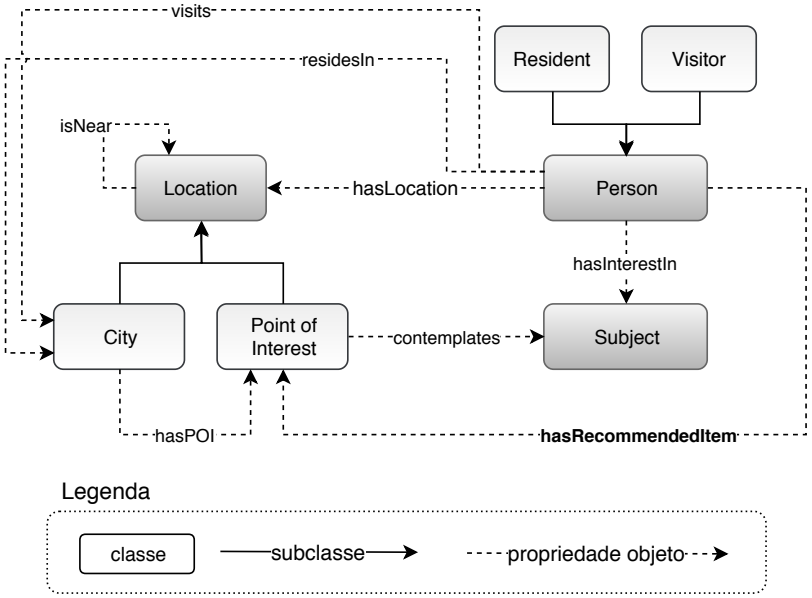
4.2.1.1 Modelo Ontológico para Cidades

A ontologia aqui descrita é definida a partir de um vocabulário comum do domínio, referenciando coleções de nomes previamente definidas (*namespaces*) da área. Foram in-

tegradas definições interpretáveis por máquina de conceitos básicos no domínio e relações entre eles, as quais expressam suas propriedades. A partir da utilização de *namespaces*, torna-se possível o compartilhamento do entendimento comum da estrutura de informação entre pessoas ou agentes de *software*, para permitir a reutilização do conhecimento de domínio, tornar explícitos os pressupostos de modelagem e, ainda, para analisar o conhecimento do domínio a que se aplica. Foram utilizados também conceitos apresentados em diferentes ontologias tais como FOAF (*Friend of a Friend*), a qual descreve pessoas (classe *Person*), suas atividades e relações na Web Semântica e OSM Semantic Network Map (2018) que expressa meios de referenciar POIs em mapeamentos colaborativos de código aberto (classe *Location*).

O modelo foi projetado e implementado em OWL (*Web Ontology Language*) com o auxílio da ferramenta Protégé[1], seguindo as diretrizes de desenvolvimento de ontologias de Noy, McGuinness et al. (2001). Como motor de inferência, para validação da consistência do modelo e de suas regras, utilizou-se o Pellet[2], o qual, de acordo com Dentler et al. (2011), apresenta grande expressividade na descrição de lógica descritiva, sendo completo, sólido, explicativo (justificação de inferências), e compatível com a ferramenta utilizada, além de possuir código aberto. A Figura 4.2 ilustra a distribuição das principais classes tratadas no modelo e algumas das interações existente entre as mesmas.

Figura 4.2: Classes Principais da Ontologia



Fonte: Acervo pessoal.

[1]https://protege.stanford.edu/
[2]https://www.w3.org/2001/sw/wiki/Pellet

4.2.2 Análise de Dados

Para o processo de análise dos dados coletados, foi utilizada uma abordagem de análise descritiva para categorizar, caracterizar, consolidar e classificar dados e convertendo-os em informações úteis. Sendo aplicados também meios de análise preditiva para identificação de padrões buscando identificar tendências futuras para o uso de tais informações. Para tanto, para a aplicação do uso de tais técnicas se deu através do uso da linguagem de programação Python, já mencionadas.

Com o intuito de analisar diferentes aspectos tais como a polaridade, confiabilidade e a influência da data de informações, que cobrem diferentes tipos de dados, são propostas então equações, as quais permeiam diferentes enfoques das análises objetivadas.

A fim de priorizar informações mais atualizadas dos recursos, a Equação 4.1 é utilizada de modo a aplicar um decaimento exponencial. Onde nessa, *diff* representa a diferença entre o dia atual e o dia em que o comentário foi realizado. Dessa forma, *reviews* recentes tendem a exercer maior influência.

$$e^{(-diff/50)} \tag{4.1}$$

Utilizou-se uma constante de decaimento exponencial (λ = 50), de modo a parametrizar o valor da distância em dias, entre o intervalo de zero a um. Dessa forma, *reviews* recentes, cuja a diferença é menor que um dia, isso é, com um intervalo de horas obtém valor um, enquanto que *reviews* com maior tempo decorrido de sua postagem tendem a se aproximar de zero. Ou seja, supondo que determinada avaliação tenha sido realizada há três dias, atribui-se a ela um valor igual a 0.94, enquanto que para outra realizada há noventa dias é atribuído o valor inferior de 0.16.

Visando analisar as polaridades expressas nos textos contidos nos comentários foram utilizados léxicos da língua portuguesa para sentimentos. O principal léxico utilizado foi o OpLexicon V3.0 disponibilizado por Souza e Vieira (2012). De modo a complementar expressões não contempladas pelo primeiro, utilizou-se ainda o léxico intitulado SentiLex, o qual é disponibilizado por Silva et al. (2010). OpLexicon foi utilizado como principal referência na análise de polaridades visto que esse possui uma versão mais atualizada quando comparada ao segundo.

A Equação 4.2 refere-se a análise de polaridade realizada sobre os *reviews*. Nessa, realiza-se o somatório das polaridades (*lex_polarity*) de um mesmo comentário extraídas

dos léxicos de sentimentos utilizados. Os dicionários de sentimentos utilizados classificam palavras positivas com valor de polaridade igual a um, negativas com valor de polaridade atribuída igual a um negativo e neutras como zero. Tal abordagem foi projetada de modo análogo a trabalhos relacionados (vide Capítulo 3), fazendo o uso de dicionários de sentimentos.

$$\sum lex\_polarity\_score \tag{4.2}$$

Tendo como exemplo a seguinte frase, é possível observar a classificação da polaridade indicada com o uso da abordagem: $\overbrace{Otimo}^{1}$ *estabelecimento! Ambiente* $\overbrace{agradavel}^{1}$ e muito $\overbrace{tranquilo}^{1}$ *área externa um pouco* $\overbrace{descuidada}^{-1}$. Assim, com o uso da Equação 4.2 o valor atribuído ao *review* analisado é de dois. Quanto maior o resultado, mais positiva a avaliação é considerada, da mesma forma tem-se que, quanto mais baixo, mais negativa essa é considerada.

Ao analisar um comentário, interessa ainda saber mais do que as informações explicitamente expressas no texto, dessa maneira, ao avaliar a confiabilidade da informação, a Equação 4.3[3] pondera a quantidade de votos positivos (*votes_up*) e negativos (*votes_down*) para todos os n comentários e ainda o número de avaliações (*n_tips*) já realizadas no site por parte do usuário que gerou o comentário i que se analisa.

$$\log_2 votes\_up_i - \log_2 votes\_down_i + \ln(n\_tips_i + 1) \tag{4.3}$$

De tal forma, um comentário que tenha recebido três votos positivos, um negativo e que tenha sido proferido por um usuário frequentemente ativo, já tendo fornecido outras dezesseis avaliações na plataforma, recebe 4.41 como resultado da aplicação da Equação 4.3. O mesmo ocorre sucessivamente então para os demais i *reviews* do estabelecimento corrente analisado, realizando a soma dos valores encontrados.

De forma a dimensionar os dados em uma escala equivalente a utilizada em redes sociais baseadas em localização, foi realizada uma padronização dos resultados gerados pelas Equações 4.1, 4.2 e 4.3 em um intervalo pré-determinado. O dimensionamento de valores é um método usado para padronizar o intervalo de variáveis independentes ou recursos de dados. No processo de tratamento de dados, essa técnica também é conhecida como normalização de dados, sendo usualmente executada durante a etapa de

[3]O logaritmo de casos de ocorrência do valor zero foram considerados como zero para fins de evitar erros matemáticos.

pré-processamento. Foi utilizado o método de normalização min-max, isso é, método que consiste em redimensionar o intervalo de recursos para dimensionar o intervalo fechado. Nesse caso, foi utilizado o intervalo restrito de um a dez, de maneira a converter os dados a um modo equivalente ao utilizado pela rede social baseada em localização analisada.

De modo a unir as variáveis previamente apresentadas (influência da data, polaridade e confiabilidade), a Equação 4.4 as considera para cada comentário *i* analisado considerando ainda variáveis parametrizáveis de peso. Essas variáveis ($\alpha, \beta e \gamma$) representam o grau de relevância de cada estimativa anteriormente calculada.

$$\sum_{i=1}^{n} polarity_i * \alpha + trust_i * \beta + decay\_date_i * \gamma \tag{4.4}$$

De modo a exemplificar, supõe-se que as variáveis referentes a polaridade, confiabilidade e relevância da data tenham sido determinadas pelas equações anteriores como sendo 8.2, 9.3 e 8.1 respectivamente. Ao aplicar a Equação 4.4 com os parâmetros da expressividade igual a um, obtém-se 25.6. Os resultados obtidos são, então, novamente normalizados em um intervalo fechado entre um e dez.

4.2.3 Recomendação de Recursos Urbanos

Esta etapa busca aplicar os conhecimentos extraídos a partir das análises realizadas, avaliando-os como possíveis indicadores a serem utilizados no processo de recomendação de recursos disponíveis. Para tanto, utilizam-se então algoritmos clássicos de sistemas de recomendação, buscando analisar os efeitos do uso de tais indicadores neste processo.

Esta etapa de recomendação dos recursos urbanos destacados na cidade de interesse pela fonte utilizada (Foursquare), é realizada através do uso da biblioteca Surprise[4]. Essa apresenta um conjunto de ferramentas em Python para construir e analisar sistemas de recomendação (filtragem colaborativa) onde diversos algoritmos são integrados, com foco na previsão de regressões. Foram então testadas variações de parâmetros de modo a buscar melhores resultados para a aplicação.

[4]http://surpriselib.com/

4.2.4 Avaliação

De modo a avaliar a capacidade de generalização dos um modelo, a partir de um conjunto de dados utilizando os indicadores de recomendação estipulados previamente, utilizou-se validação cruzada. De acordo com Kohavi et al. (1995), na validação cruzada de *k-fold*, o conjunto de dados D é dividido aleatoriamente em k subconjuntos mutuamente exclusivos de tamanho aproximadamente igual. O indutor é treinado e testado k vezes, de modo que a cada vez um subconjunto é utilizado para teste e os $k - 1$ restantes são utilizados para aferir os parâmetros e então calcula-se a acurácia do modelo. Dessa forma, segundo esse autor, a técnica estima o número total de predições corretas, dividido pelo número de instâncias no conjunto de dados.

Tal técnica é empregada em problemas onde o objetivo é a predição, de modo que se busca estimar o quão preciso é esse modelo a um novo conjunto de dados. Em um estudo realizado por Rodriguez, Perez e Lozano (2010), após realizar uma análise de sensibilidade da validação cruzada k-*fold* na estimativa de erro de predições, recomenda-se o uso de $k = 5$ ou $k = 10$, dado o baixo viés apresentado com o uso de tais valores. Dessa forma, ao aplicar esta abordagem, foram utilizados tais valores. Com o uso dessa técnica foi avaliado o erro médio absoluto (do inglês Mean Absolute Error – MAE) e da raiz do erro médio quadrático (do inglês Root Mean Squared Error – RMSE). Essas são utilizadas para medir a acurácia de variáveis contínuas, sendo de acordo com Yang et al. (2014) também amplamente utilizadas para avaliação em sistemas de recomendação.

O erro médio absoluto mede a magnitude média dos erros em um conjunto de previsões, sem considerar sua direção. É a média sobre a amostra de teste das diferenças absolutas entre a previsão e a observação real, em que todas as diferenças individuais possuem o mesmo peso. O RMSE é uma regra de pontuação quadrática que também mede a magnitude média do erro, entretanto essa é a raiz quadrada da média das diferenças quadradas entre previsão e observação real feita.

Assumir a raiz quadrada dos erros quadrados médios possui implicações relevantes. Dado que os erros são elevados antes da média, o RMSE atribui um peso relativamente alto a erros grandes, significando assim que o RMSE é útil quando erros grandes são particularmente indesejáveis. Dessa forma, neste trabalho são consideradas ambas as métricas de modo a analisar diferentes aspectos referentes a acurácia obtida.

De modo a comparar os resultados obtidos aos do uso de um recomendador que utilize a nota como critério principal de avaliação, é realizado então um teste de Wilcoxon

sobre os resultados. Esse é um teste de hipóteses não paramétrico utilizado quando se deseja comparar duas amostras relacionadas, emparelhadas ou medidas repetidas em uma única amostra para avaliar se as medidas de posição dessas são iguais no caso em que as amostras são dependentes. Esse tipo de teste costuma ser utilizado em amostras de métricas de avaliação de modelos para averiguar se a diferença de habilidade entre os modelos de aprendizado de máquina é significativa.

4.3 Algoritmos

Nesta seção são descritos os algoritmos que têm sua execução proposta como teste sobre o conjunto de dados estudado neste trabalho. A escolha por eles se deu pois, de acordo com autores como Amatriain e Pujol (2015) e Adomavicius e Tuzhilin (2011), representam algumas das técnicas mais tradicionais de recomendação. Todas as versões dos algoritmos utilizados são implementadas através da biblioteca Surprise em sua versão 1.0.6.

4.3.1 Decomposição de valores singulares

De acordo com Amatriain e Pujol (2015), o algoritmo de Decomposição de Valor Singular (do inglês, *Singular Value Decomposition* – SVD) apresenta uma técnica poderosa para redução de dimensionalidade, onde essa é uma realização particular da abordagem da Fatoração de Matrizes. O objetivo de uma decomposição de SVD é encontrar um espaço de caractere dimensional menor, em que os novos recursos representam "conceitos" e a intensidade de cada conceito no contexto da coleção é computável.

A recomendação r_{ui} realizada pelo algoritmo é dada pela Equação 4.5.

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u \tag{4.5}$$

Caso o usuário u seja desconhecido, assume-se que o bias b_u e os fatores $p_u$são nulos (zero). O mesmo é aplicado para o item i com b_i e q_i.

Foi utilizada ainda uma variação desse algoritmo, também implementada na biblioteca de recomendação utilizada, identificado como SVD++. Essa extensão do algoritmo tradicional considera em seu processamento avaliações implícitas. No algoritmo SVD++

a recomendação r_{ui} realizada é dada pela Equação 4.6.

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u (p_u + |I_u|^{1/2} \sum_{j \in I_u} y_j) \tag{4.6}$$

Onde os termos y_j são um novo conjunto de fatores de itens que capturam predições implícitas. Uma predição implícita descreve o fato de um usuário ter classificado um item j, independentemente do valor predito.

Assim como no caso tradicional, em casos onde o usuário u é desconhecido, assume-se que o bias b_u e os fatores p_u são nulos (zero). O mesmo é aplicado para o item i com b_i e q_i.

Os algoritmos implementados na biblioteca utilizada (*surprise*), são parametrizáveis. Desse modo, estes permitem variações tais como alterações o número de fatores utilizados, fator de regularização, taxa de aprendizado, entre outros.

4.3.2 Vizinhos mais próximos

Os classificadores baseados em instâncias funcionam armazenando registros de treinamento e usando-os para prever o rótulo de classe de casos não vistos. Dado um ponto a ser classificado, segundo Amatriain e Pujol (2015), o classificador k-NN encontra os k pontos mais próximos (vizinhos mais próximos) dos registros de treinamento. E, logo, atribui o rótulo de classe de acordo com os rótulos de classe de seus vizinhos mais próximos. A ideia subjacente é que, se um registro pertence a uma vizinhança em particular, onde um rótulo de classe é predominante, é porque o registro provavelmente pertence a essa mesma classe.

De acordo com Gama et al. (2011), na execução desse algoritmo então são obtidos os k vizinhos, de modo que cada vizinho vota então em uma classe. Desse modo, as previsões de diferentes vizinhos são agregadas de forma a classificar o ponto de teste. Em Amatriain e Pujol (2015), afirma-se que essa é uma das abordagens mais comuns na técnica de filtragem colaborativa. Assim, o conjunto de predição do algoritmo é dado pelas Equações 4.7 e 4.8, para recomendações baseadas em usuário e item respectivamente.

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i^k(u)} sim(u,v) \cdot r_{vi}}{\sum_{v \in N_i^k(u) sim(u,v)}} \tag{4.7}$$

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i^k(i)} sim(i,j) \cdot r_{uj}}{\sum_{v \in N_i^k(j) sim(i,j)}} \tag{4.8}$$

Utilizou-se também uma variação do algoritmo tradicional de vizinho mais próximo, o k-NN *Baseline*. Esse implementa um algoritmo básico de filtragem colaborativa que leva em conta uma predição de linha de base. O conjunto de predição desse algoritmo é dado pelas Equações 4.9 e 4.10, para recomendações baseadas em usuário e item respectivamente.

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{v \in N_i^k(u)} sim(u,v) \cdot (r_{vi} - bvi)}{\sum_{v \in N_i^k(u)} sim(u,v)} \tag{4.9}$$

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{v \in N_u^k(i)} sim(i,j) \cdot (r_{uj} - buj)}{\sum_{v \in N_u^k(j)} sim(i,j)} \tag{4.10}$$

Os algoritmos implementados na biblioteca utilizada (*surprise*) são parametrizáveis, permitindo variações tais como alterações no valor de k, valor mínimo de k para agregação, medidas de similaridade e estimativas de *baseline*.

4.4 Execução

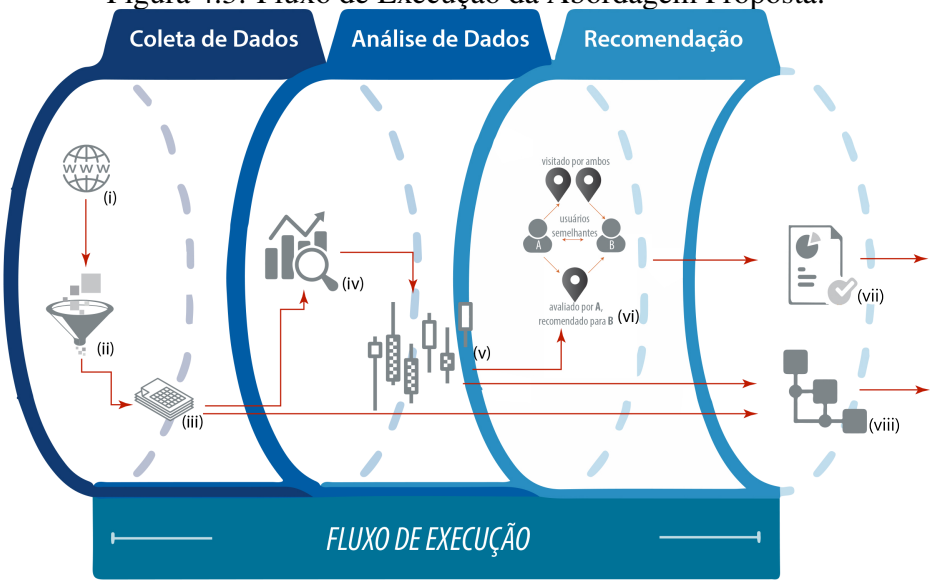
Ao final da concepção da proposta realizada, seguindo a estrutura determinada, tem-se um fluxo de execução tal como exemplificado na Figura 4.3. Em um primeiro estágio da execução (coleta de dados) do *pipeline* tem-se a seleção de uma fonte de dados de LBSNs *(i)* e a coleta *(ii)* destes desenvolvida de modo *ad-hoc*. A partir disso, os dados são então organizados e tratados (pré-processados) *(iii)*.

No estágio de análise de dados *(iv)*, os dados obtidos anteriormente são então analisados de forma a extrair informações relevantes. Para tanto, são empregados meios de descoberta de conhecimentos em bases textuais propostos para a obtenção de *features* de polaridade. São observados ainda dados de usuários a fim de obter informações referentes a idoneidade das informações providas por esse. Dados relativos a data de informações são também considerados visando prover meios de priorizar *reviews* recentes em detrimento dos antigos. A partir da observação dos dados analisados são extraídas as *features* de interesse (polaridade, confiabilidade e data), gerando os chamados indicadores de recomendação *(v)* a serem utilizados para a estimativa de avaliação de cada ponto de interesse.

Já na fase de recomendação *(vi)* são utilizados algoritmos de abordagens de filtragem colaborativa implementados com o uso de uma biblioteca externa auxiliar. Os algoritmos são executados seguindo o padrão de parâmetros: *user item rating*. Como *rating* foi então utilizada a estimativa gerada a partir da aplicação dos indicadores gerados no estágio anterior, buscando verificar a eficiência de seu uso no processo de recomendação de recursos urbanos.

É realizada ainda a aplicação dos algoritmos de recomendação empregando as avaliações padrão utilizadas em redes sociais baseadas em localização, para fins de comparação com a proposta sugerida. A nota atribuída[5] aos estabelecimentos pelo Foursquare, considera *likes*, *dislikes* e popularidade de cada local. O intervalo de notas utilizado no site varia em uma escala crescente de um até dez, sendo dez a nota de maior estimativa atingida. As execuções foram realizadas utilizando validação cruzada k-*fold*, variando o valor de k em cinco e dez. Por fim, é realizada uma avaliação *(vii)* dos resultados utilizando o teste estatístico de Wilcoxon de modo a verificar a diferença de uso de cada estimativa utilizada. Além da avaliação gerada, como saída, tem-se ainda como modelo de representação do conhecimento gerado a ontologia *(viii)* populada com o uso de informações advindas dos indicadores e dos demais dados estudados.

Figura 4.3: Fluxo de Execução da Abordagem Proposta.



(i) Seleção da fonte de dados; *(ii)* Coleta de dados; *(iii)* Pré-processamento; *(iv)* Análise de dados; *(v)* Indicadores de recomendação; *(vi)* Recomendação de recursos urbanos; *(vii)* Avaliação; *(viii)* Ontologia populada a partir dos dados extraídos e gerados.

Fonte: Acervo pessoal.

[5]https://support.foursquare.com/hc/en-us/articles/201109274-Place-ratings

Como demonstrado, o fluxo de execução utilizado decorre da organização da proposta previamente estruturada. Todos os estágios foram implementados em Python e pensados de modo a estabelecer uma conexão entre os diferentes módulos. Cabe salientar que os códigos-fonte encontram-se disponíveis na plataforma GitHub[6], mais especificamente em: <https://brendasalenave.github.io/dissertacao/>

[6]Plataforma de hospedagem de código-fonte com controle de versão utilizando Git.

5 EXTRAÇÃO DE INDICADORES DE RECOMENDAÇÃO A PARTIR DE DADOS ESTRUTURADOS E NÃO-ESTRUTURADOS

Neste Capítulo são descritas as aplicações das abordagens de análise descritas no Capítulo 4. Para tanto, este encontra-se organizado do seguinte modo: a Seção 5.1 apresenta a descrição do conjunto de dados utilizado para fins de realização das análises de testes sobre os indicadores gerados a partir dos dados dessas; a Seção 5.2 apresenta as análises realizadas para extração de *features* auxiliares ao processo de recomendação de recursos urbanos; por fim, a Seção 5.3 apresenta o modelo utilizado para representação do conhecimento capturado.

5.1 Conjunto de Dados

Como fonte de dados para estudos, adotou-se a rede Foursquare, cenário o qual, de acordo com Gao et al. (2015), apresenta uma das mais populares redes sociais baseadas em localização. Dessa maneira, optou-se por analisar dados provenientes da cidade de Gramado, no Rio Grande do Sul (IBGE, 2019a), como objeto de estudos uma vez que essa é considerada pelo TripAdvisor como um dos principais destinos turísticos de inverno do país. Foram coletados dados disponíveis até 20 de dezembro de 2018.

Assim, a partir da implementação do coletor de informações descrito na Seção 4.2.1 foram obtidas as informações descritas na Tabela 5.1. Essa apresenta as variáveis obtidas juntamente com a cardinalidade (valores distintos) e o pré-processamento realizado para cada. As Figuras 5.1 e 5.2 apresentam ainda descrições adicionais realizadas sobre o conjunto de dados estudados de LBSNs.

A Figura 5.1 apresenta uma visão geral do conjunto de dados inicialmente extraído, gerada com o auxílio da biblioteca *Pandas Profiling*[1]. Nessa é possível observar que o conjunto possui um total de 13 informações distintas sobre cada instância, sendo 7 classificadas como numéricas e 6 como categóricas. A Figura mostra, ainda, avisos referentes a distribuição e cardinalidade dos dados.

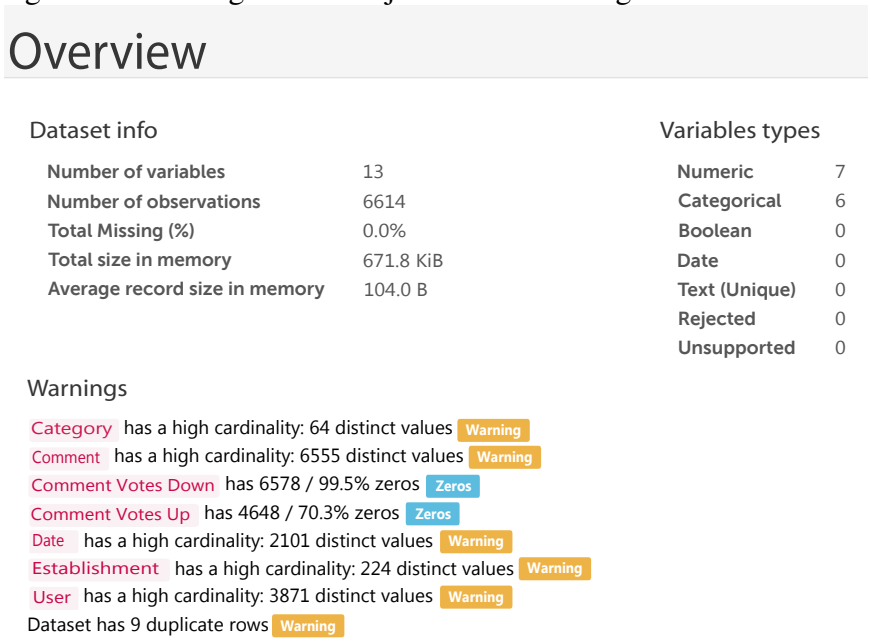
Já a Figura 5.2 apresenta a visão geral do conjunto, após esse ser pré-processado utilizando as abordagens especificadas na Tabela 5.1. Esta etapa de tratamento dos dados teve como principal alvo os comentários fornecidos pelos usuários. Todos os comentários

[1] https://pypi.org/project/pandas-profiling/

Tabela 5.1: Conjunto variáveis coletadas

Variável	Cardinalidade	Pré-processamento
Nome do Estabelecimento	24	Conversão Simbólico-Numérico
Categoria	64	Conversão Simbólico-Numérico
Média Avaliações	39	Nenhum
Total Avaliações	164	Nenhum
Latitude	222	Nenhum
Longitude	222	Nenhum
Comentário	6490	Remoção de *stopwords* Tokenização Lematização
Data Comentário	2101	Formatação
Votos Favoráveis	21	Nenhum
Votos Contrários	3	Nenhum
Usuário	3868	Conversão Simbólico-Numérico
Gênero Usuário	3	Removido
Total Dicas Usuário	305	Nenhum

Figura 5.1: Visão geral do conjunto de dados originalmente coletado



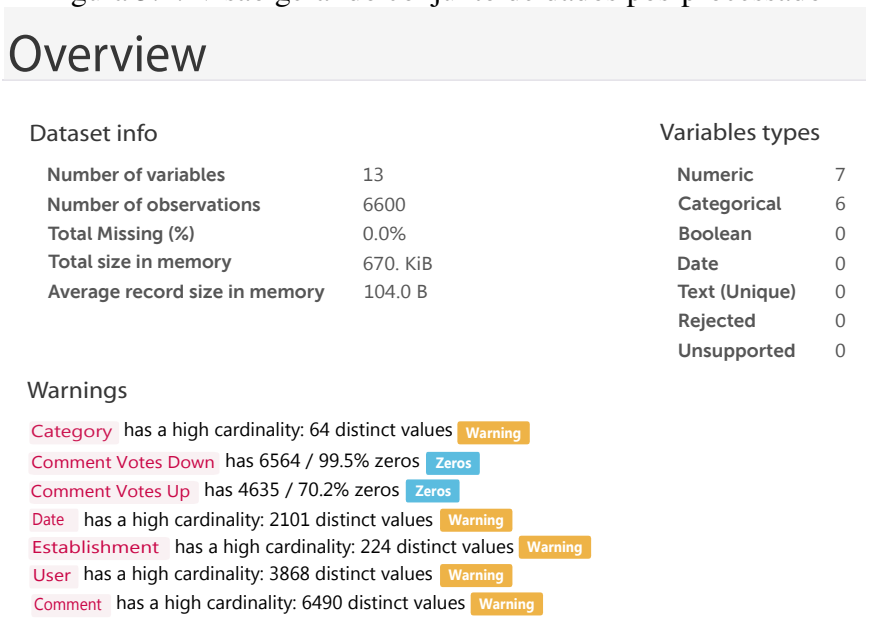
Fonte: Acervo pessoal.

foram submetidos ao processo de remoção de palavras vazias (*stopwords*). Como tokenizador foi utilizado o RegexTokenizer[2], separando todas as palavras individualmente. Para melhor uniformizar o conteúdo de cada avaliação (comentário), essas foram ainda lematizadas com o uso da biblioteca Spacy[3], de maneira a deflexionar uma palavra para determinar o seu lema.

[2]https://www.nltk.org/_modules/nltk/tokenize/regexp.html
[3]https://spacy.io/

Ainda na etapa de pré-processamento, foi realizada a padronização das datas de realização de cada comentário convertendo-as para o padrão: $dd/mm/aaaa$. Para as variáveis de referência ao nome do estabelecimento, categoria e nome de usuário foram aplicadas transformações de dados, realizando conversões simbólico-numéricas de modo a anonimizar usuários e facilitar o processamento dos demais dados. A variável de indicação do gênero do usuário foi descartada por interpretar-se uma baixa relevância para o processo.

Figura 5.2: Visão geral do conjunto de dados pós-processado



Fonte: Acervo pessoal.

Já como fonte de dados referentes a cidades de forma geral, utilizou-se a plataforma DataViva a fim de coletar informações oficiais sobre diferentes tipos de informações de municípios brasileiros. Foram extraídos dados tais como descrição do município, localização geográfica, bem como outras referentes a aspectos sociais e econômicos, visando assim complementar o sistema, tornando-o mais robusto e informativo.

A Tabela 5.2 apresenta o conjunto final de variáveis utilizadas e as respectivas fontes de coleta. Cabe ressaltar que foram utilizadas variáveis comuns a diferentes LBSNs e disponíveis no Foursquare, visando permitir sua aplicação a conjuntos de dados advindos de diferentes bases. As variáveis extraídas do DataViva foram agregadas de modo que, ao visualizar a composição geral de um conjunto de dados de uma determinada cidade, possa-se observar e inferir sobre o diferentes o domínio sob diferentes perspectivas, permitindo diferentes análises e observações. Todas as variáveis apresentadas na tabela

foram utilizadas na composição da ontologia utilizada para representação dos dados.

Tabela 5.2: Conjunto de variáveis finais utilizadas.

Variável	Fonte
Comentário	Foursquare
Data Comentário	Foursquare
Expectativa de Vida	Data Viva
Latitude	Foursquare
Longitude	Foursquare
Média Avaliações	Foursquare
Nome do Estabelecimento	Foursquare
Perfil Geral	Data Viva
População Total	Data Viva
Total Avaliações	Foursquare
Total Dicas Usuário	Foursquare
Usuário	Foursquare
Votos Contrários	Foursquare
Votos Favoráveis	Foursquare

5.2 Indicadores de Recomendação

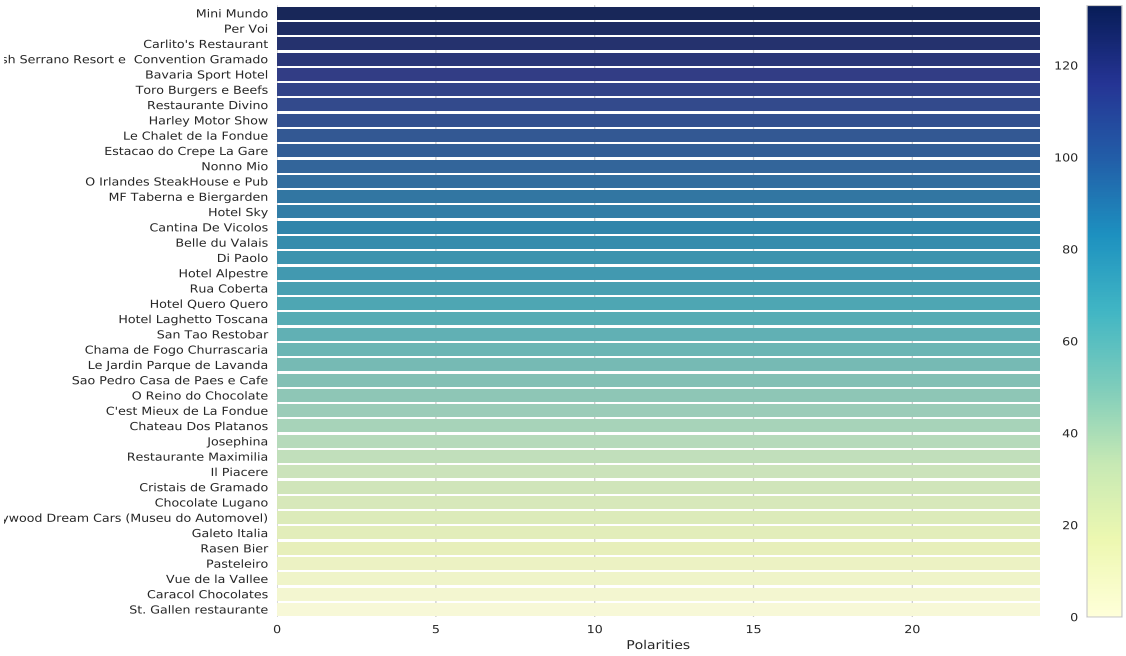
Com foco na extração de *features* para análise de dados provenientes da Web, foram empregadas as abordagens descritas na Seção 4.2.2. Assim, buscou-se explorar informações consideradas relevantes para fins de levantamento de características intrínsecas presentes nos dados estudados (descritos na Seção 5.1).

Como explicitado na Seção 4.2.2, para realização da análise de polaridade dos *reviews*, foram utilizados dois léxicos de sentimentos. Cada palavra lematizada foi submetida à etapa de análise, buscando identificar sua valência para assim ser classificada como positiva ou negativa. Desse modo, seguindo a Equação 4.2, tem-se a polaridade de cada comentário e consequentemente de cada estabelecimento avaliado.

As Figuras 5.3 e 5.4 apresentam as polaridades positivas e negativas, respectivamente, acumuladas para cada estabelecimento analisado. Para fins de visualização, as figuras apresentam somente os primeiros 40 itens, ordenados de forma decrescente pela quantidade de polaridade (positiva ou negativa) reunida, onde essa é representada pela tonalidade de cores.

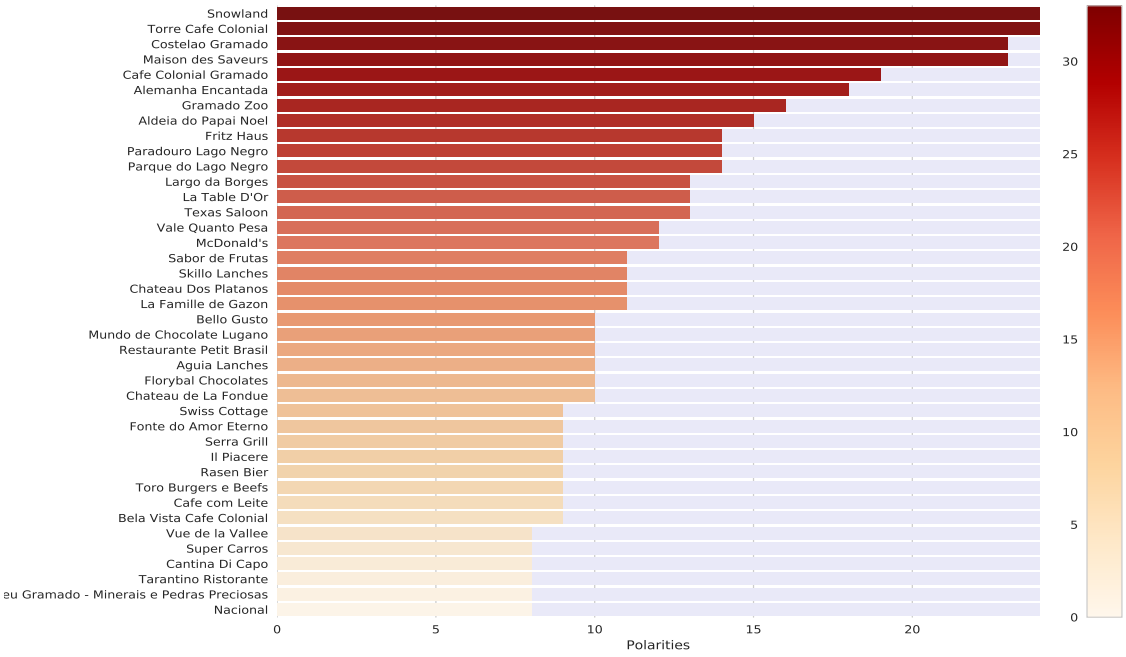
Utilizando Equação 4.3, tem-se a confiabilidade de cada *review* coletado por estabelecimento, e assim consequentemente a confiabilidade acumulada para cada um destes. Para fins de visualização, a figura apresenta somente os primeiros 40 itens, ordenados de

Figura 5.3: Polaridades Positivas



Fonte: Acervo pessoal.

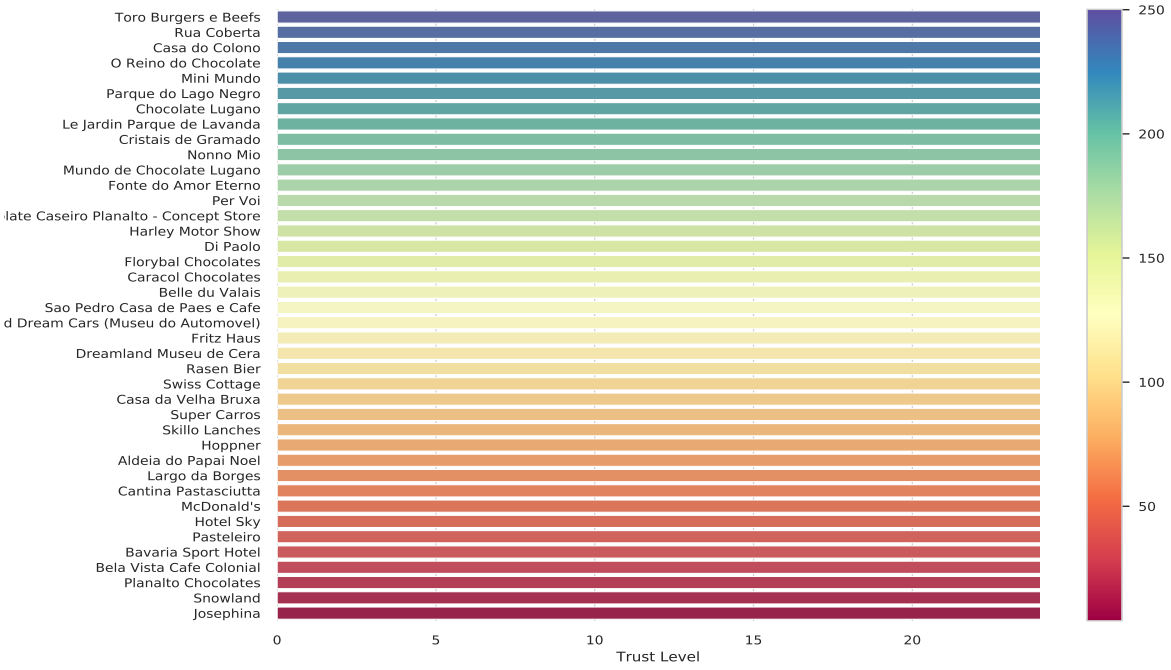
Figura 5.4: Polaridades Negativas



Fonte: Acervo pessoal.

forma decrescente pela confiabilidade reunida de cada comentário, onde a esta é representada pela tonalidade de cores.

Figura 5.5: Confiabilidade

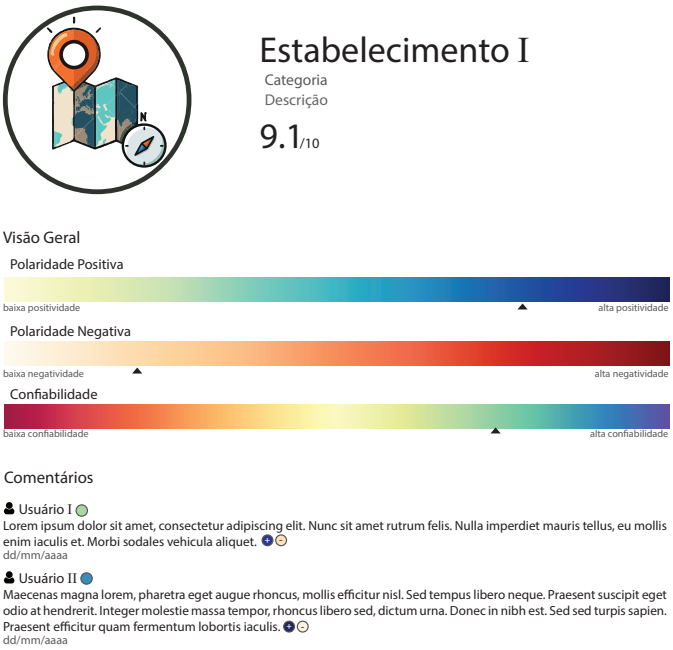


Fonte: Acervo pessoal.

Ao utilizar a Equação 4.4, realiza-se então a combinação das *features* anteriores (polaridades e confiabilidade) a Equação 4.1 aplicando a cada comentário um fator de decaimento sobre a data em que foi realizado, priorizando dessa forma as avaliações mais recentes. De tal forma, tem-se a integração de diferentes tipos de informações advindas de dados extraídos a partir de fontes estruturadas e não-estruturadas, de modo a compreender diferentes aspectos da informação.

A Figura 5.6 apresenta um exemplo de representação de recomendação que pode ser gerada a partir do uso de tais *features* e indicadores de avaliação. Nessa, destacam-se as características gerais encontradas na análise de dados de determinado estabelecimento. Ela evidencia ainda a idoneidade de usuários que realizaram as avaliações e as respectivas polaridades dessas, que resultaram na estimativa de *rating* gerado.

Figura 5.6: Exemplo de Aplicação



Fonte: Acervo pessoal.

5.3 Ontologia

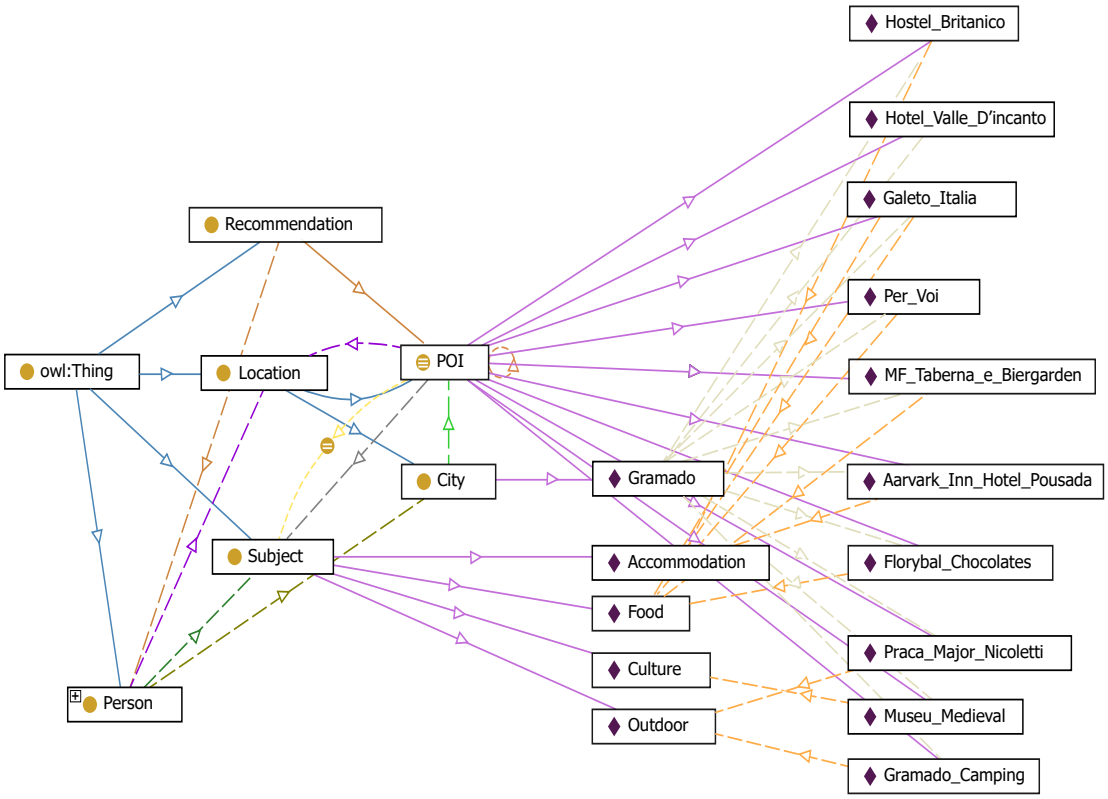
Segundo Isotani e Bittencourt (2015), o uso de ontologias tem-se mostrado uma tecnologia chave na criação de aplicações mais adequadas a lidar com grandes quantidades de informações de modo inteligente. Dessa forma, visando contribuir para o desenvolvimento de um modelo ontológico voltado ao domínio de cidades inteligentes, foi implementado um método para população automatizada da ontologia base descrita na Seção 4.2.1.1.

A Figura 5.7 apresenta uma versão simplificada do resultado da população do modelo de representação utilizado, a partir de uma pequena amostra do conjunto. Tal Figura foi gerada com o auxílio da ferramenta OntoGraf[4], a qual oferece suporte de navegação interativa pelos relacionamentos das ontologias de OWL.

Como demonstrado no Capítulo 3, o uso de ontologias como suporte a sistemas de recomendação é tido como um elemento de grande importância para o desenvolvimento da pesquisa. Tem-se ainda, em Bagherifard et al. (2017), o desenvolvimento de estudos voltados ao aprimoramento de acurácia em algoritmos de recomendação baseados em filtragem colaborativa com suporte em ontologias. Dessa forma, o demonstra esta abordagem como forma de atingir melhores resultados utilizando algoritmos baseados em

[4]https://protegewiki.stanford.edu/wiki/OntoGraf

Figura 5.7: Visão simplificada da ontologia populada a partir do conjunto de dados estudado.



Fonte: Acervo pessoal.

vizinhança tal como K-NN.

Com base na ontologia utilizada (Seção 4.2.1.1) populada com o conjunto de dados de interesse, desconsiderando variáveis removidas no estágio de pré-processamento, torna-se possível em trabalhos fututos então o desenvolvimento de sistemas de recomendação fundamentados em tal tipo de representação do conhecimento.

Assim, tem-se ainda a possibilidade de integração da ontologia à aplicação do uso de indicadores de recomendação, visando a projeção de um sistema mais robusto e acurado.

6 APLICAÇÃO DE INDICADORES NO PROCESSO DE RECOMENDAÇÃO DE RECURSOS URBANOS

Neste Capítulo são descritos os experimentos empreendidos a fim de realizar a aplicação das métricas geradas a partir do processo de análise de dados em algoritmos de recomendação. Para isso utilizou-se os indicadores de recomendação extraídos e apresentados no Capítulo 5. Assim, a Seção 6.1 apresenta os resultados alcançados a partir de experimentos realizados tendo como foco as métricas de acurácia atingidas com o uso dos indicadores propostos, juntamente com uma breve discussão a cerca destes resultados.

6.1 Resultados e Discussão

De forma a validar o uso dos indicadores propostos no processo de recomendação de recursos urbanos, realizou-se a aplicação da métrica constituída a partir do uso das *features* previamente geradas em algoritmos tradicionais de recomendação. A execução dos algoritmos também se deu utilizando a avaliação geral (*rating*) dos estabelecimentos no ambiente analisado.

Utilizando a biblioteca *Surprise*, realizou-se a aplicação dos algoritmos descritos na Seção 4.3 com suas respectivas variações. De modo a atingir melhores resultados, foi realizada uma sequência de testes sobre os algoritmos para cada métrica. Foi então utilizada uma funcionalidade da biblioteca, chamada *GridSearchCV*, a qual busca responder qual combinação de parâmetros produz melhores resultados. A execução de cada algoritmo e suas variações se deu então utilizando a combinação de parâmetros gerada pela ferramenta. Foram gerados histogramas (Figuras 6.1, 6.2, 6.3, 6.4, 6.5 e 6.6) para demonstrar quão diferentes são as previsões com ambas as métricas, com a contagem do número de previsões para cada valor predito e utilizando as melhores configurações indicadas pelo uso do *GridSearchCV*. De tal forma, a quantidade de predições fornecidas é representada no eixo Y, enquanto que os valores de *rating* são denotados no eixo X. Em todas as figuras sempre é mostrada a abordagem tradicional (à esquerda) e a proposta (à direita).

Visando explorar os melhores resultados com a execução dos algoritmos SVD e SVD++, foram então testadas as variações de parâmetros descritas na Tabela 6.1. Para ambos os algoritmos, os valores finais utilizados foram os que estão em destaque na tabela

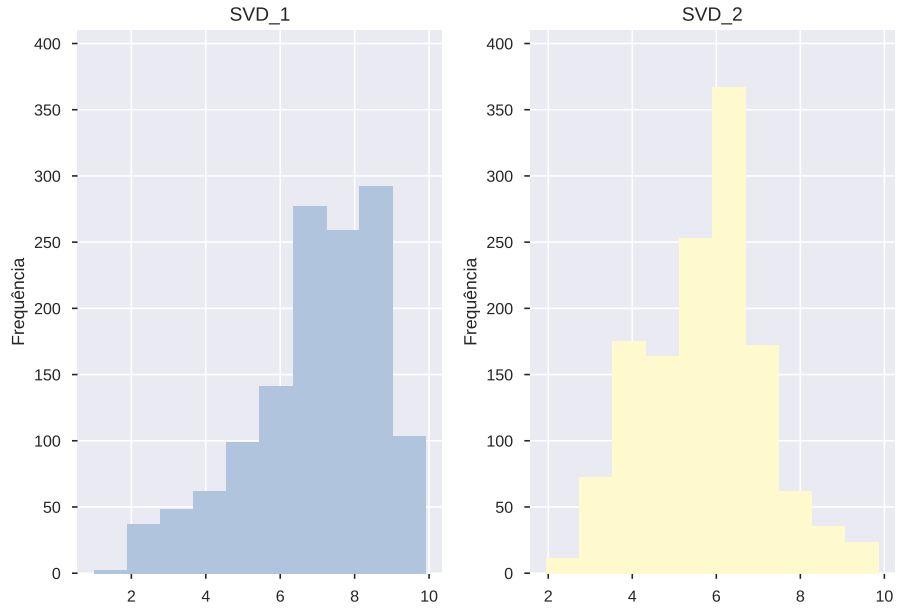
apresentada a seguir.

Tabela 6.1: Variação de parâmetros dos algoritmos SVD e SVD++.

Parâmetro	Valores Testados	Descrição
n_epochs	[5, 10, 20, **50**]	Número de iterações do procedimento SGD
lr_all	[0.002, **0.005**]	Taxa de aprendizado para todos os parâmetros
reg_all	[**0.005**, 0.02, 0.4, 0.6]	Termo de regularização para todos os parâmetros

A Figura 6.1 apresenta o histograma gerado para as predições realizadas utilizando o algoritmo SVD. Nela é possível perceber a diferença de predição ao se utilizar o algoritmo com cada métrica. Utilizando a abordagem tradicional, a maior taxa de predições do algoritmo concentra-se nos *ratings* mais altos, enquanto que com o uso a métrica proposta ocorre o inverso.

Figura 6.1: Histograma de previsões para cada valor de predição utilizando SVD

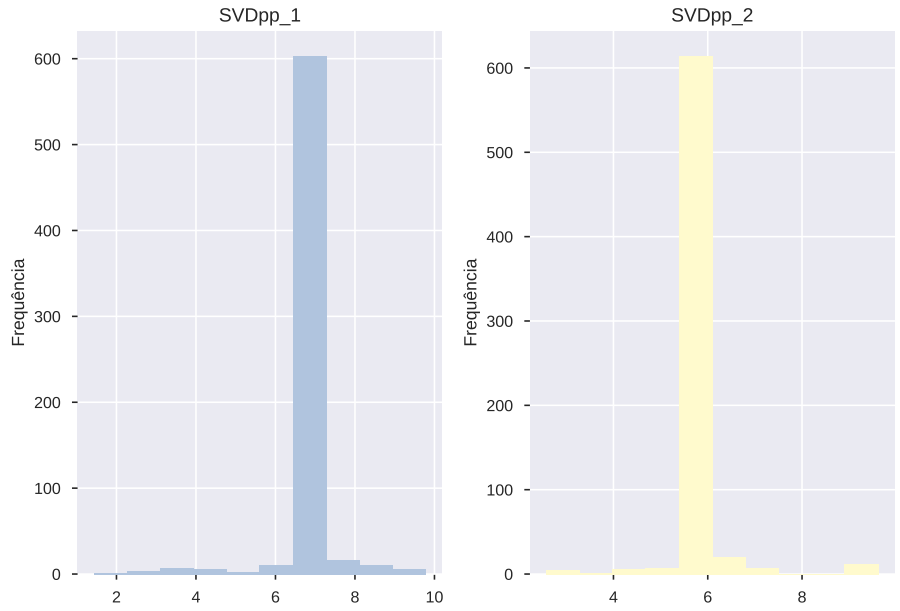


Fonte: Acervo pessoal.

Já o histograma gerado a partir das predições utilizando o algoritmo SVD++, apresentado na Figura 6.2, ilustra as predições concentradas em algumas notas. Da mesma forma, o uso da avaliação padrão mantém-se com maior número de predições para notas mais altas.

Todos os valores de parâmetros utilizados foram extraídos através de testes realizados com a aplicação da funcionalidade de *GridSearchCV* da biblioteca *Surprise*. Desta forma, para execução dos algoritmos $K - NN$ baseados em usuário, foram testadas as variações de parâmetros descritas na Tabela 6.2. Ao considerar a abordagem baseada em

Figura 6.2: Histograma de previsões para cada valor de predição utilizando SVD++



Fonte: Acervo pessoal.

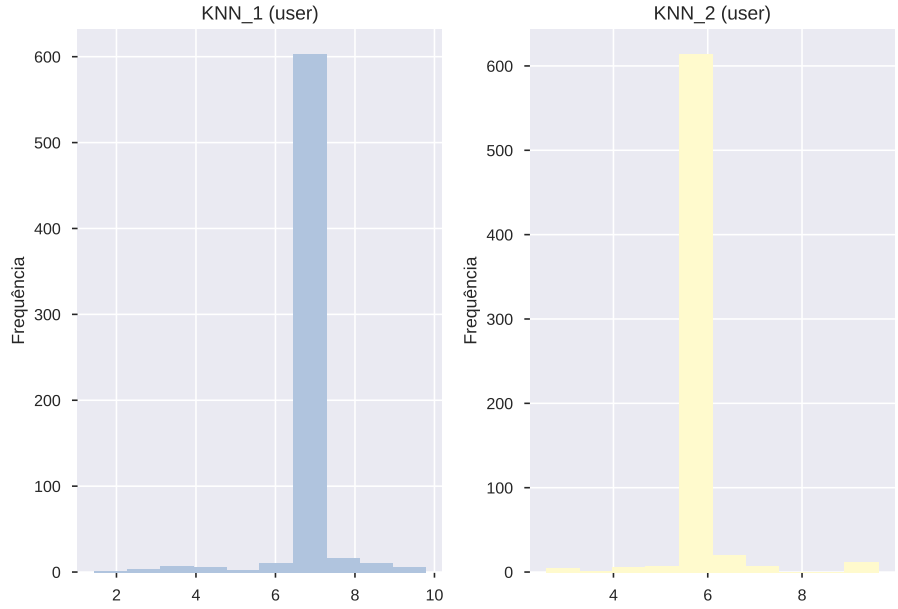
usuário para ambas as métricas, foi utilizado um valor de $k = 5$, e *msd* como medida de similaridade com um suporte mínimo de 5. O número mínimo de vizinhos para agregação manteve-se em 1.

Assim como no caso anterior (SVD++), o histograma gerado a partir das predições do algoritmo $K - NN$ baseado em item mostra-se concentrado em poucas faixas de avaliação, tal como apresentado na Figura 6.3.

Da mesma forma, os valores de parâmetros utilizados para os algoritmos $K - NN$ baseados em item e em usuário, foram extraídos através de testes realizados com a aplicação da funcionalidade de *GridSearchCV* da biblioteca de recomendação utilizada. Desta maneira, para estes algoritmos ($K - NN$ baseados em item e em usuário), foram testadas as variações de parâmetros descritas na Tabela 6.2. Dessa maneira, para a métrica-padrão de *rating* foi utilizado um valor de $k = 100$, e *cosine* como medida de similaridade com um suporte mínimo de 1. O parâmetro de encolhimento (*shrinkage*) na similaridade não foi utilizado (0), esse auxilia a evitar *overfitting* quando apenas poucas predições prévias estão disponíveis. Para a métrica proposta utilizou-se $k = 5$ e *msd* como medida de similaridade com um suporte mínimo de 1. O número mínimo de vizinhos para agregação manteve-se 1 para ambas as métricas consideradas.

A Figura 6.4 apresenta o histograma gerado a partir das predições do algoritmo $K - NN$ baseado em usuário, essa se mostra distribuída de modo bastante semelhante a abordagem do algoritmo baseada em item (Figura 6.3). Do mesmo modo, as predições

Figura 6.3: Histograma de previsões para cada valor de predição utilizando K-NN



Fonte: Acervo pessoal.

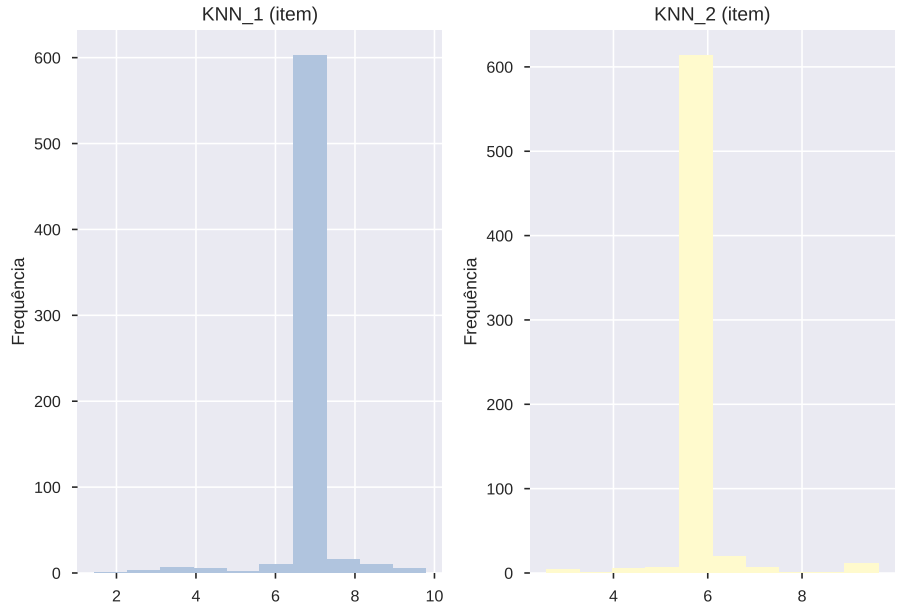
Tabela 6.2: Variação de parâmetros do algoritmo K-NN.

Parâmetro	Valores Testados	Descrição
k	[5, 10, 20, 40, 100]	Número (máximo) de vizinhos a serem considerados para agregação
min_k	[1, 5, 10]	Número mínimo de vizinhos a considerar para agregação
sim_options	name: [cosine, msd] min_suport: [1, 5] shrinkage: [0, 50, 100]	Opções para a medida de similaridade

realizadas utilizando a abordagem proposta (à direita da figura) aglomeram-se em estimativas de notas mais baixas do que as previstas pela avaliação utilizada pelo site (à esquerda da figura).

Para execução dos algoritmos K-NN Baseline foram testadas as variações de parâmetros descritas na Tabela 6.3. Ao considerar a abordagem baseada em usuário, foi utilizado um valor de $k = 100$ para a métrica proposta, tendo *pearson_baseline* como medida de similaridade com um suporte mínimo de 1 e com a taxa de aprendizado igual à 0.1. Como *baseline* foi utilizado o método *sgd*, com uma taxa de aprendizado igual à 0.1, com 50 iterações (*epochs*) e fator de regularização igual à 0.02. O número mínimo de vizinhos para agregação manteve-se 1. Já para a métrica tradicional, os parâmetros mantiveram-se alterando apenas $k = 5$ com uma taxa de aprendizado de 0.00005. Em tais casos, os valores de parâmetros utilizados foram também obtidos através de testes realizados com a aplicação da funcionalidade de *GridSearchCV* da biblioteca *Surprise*.

Figura 6.4: Histograma de previsões para cada valor de predição utilizando K-NN baseado em item



Fonte: Acervo pessoal.

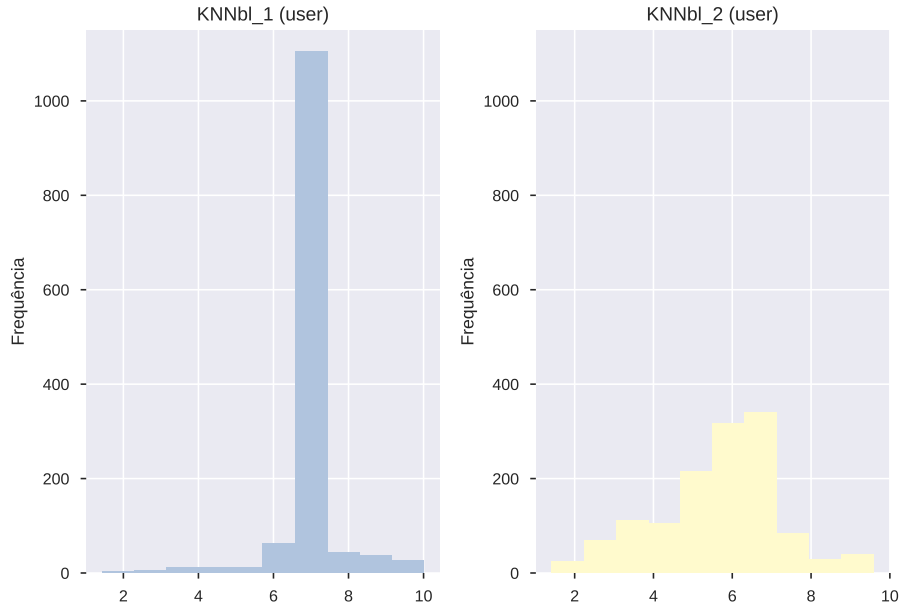
Ao aplicar tais parâmetros no algoritmo K-NN Baseline baseado em usuário, a distribuição das predições fornecidas por este mostram-se bastante distintas, tal como retratado na Figura 6.5, ao variar a estimativa de *rating*. Enquanto que o uso da avaliação padrão aponta para previsões condensadas em uma faixa restrita de nota, o uso da abordagem proposta, apresentado à direita da figura, mostra-se distribuído, ou seja, indicando predições para diferentes atribuições de nota.

Já com a utilização do algoritmo K-NN Baseline baseado em item, a melhor configuração encontrada consistiu no uso de *pearson_baseline* como medida de similaridade com um suporte mínimo de 1. O parâmetro de encolhimento (*shrinkage*) utilizado na similaridade foi igual a 100. O valor de k foi configurado como 5, e seu valor mínimo (*k_min*) como sendo 1.

Ao variar a abordagem de aplicação do algoritmo K-NN Baseline, empregando-o de modo baseado em item seguindo os parâmetros indicados, a distribuição das predições fornecidas por esse mostram-se distintas, tal como retratado na Figura 6.6, ao variar a estimativa de *rating*. Enquanto que o uso da avaliação padrão aponta para previsões condensadas em uma faixa restrita de nota, assim como na aplicação da abordagem anterior, o uso da estimativa proposta, apresentado à direita da figura, se mostra distribuído, ou seja indicando predições para diferentes atribuições de nota.

Os histogramas mostrados nas Figuras 6.1, 6.2, 6.3, 6.4, 6.5 e 6.6) apresentam ca-

Figura 6.5: Histograma de previsões para cada valor de predição utilizando K-NN Baseline baseado em usuário.



Fonte: Acervo pessoal.

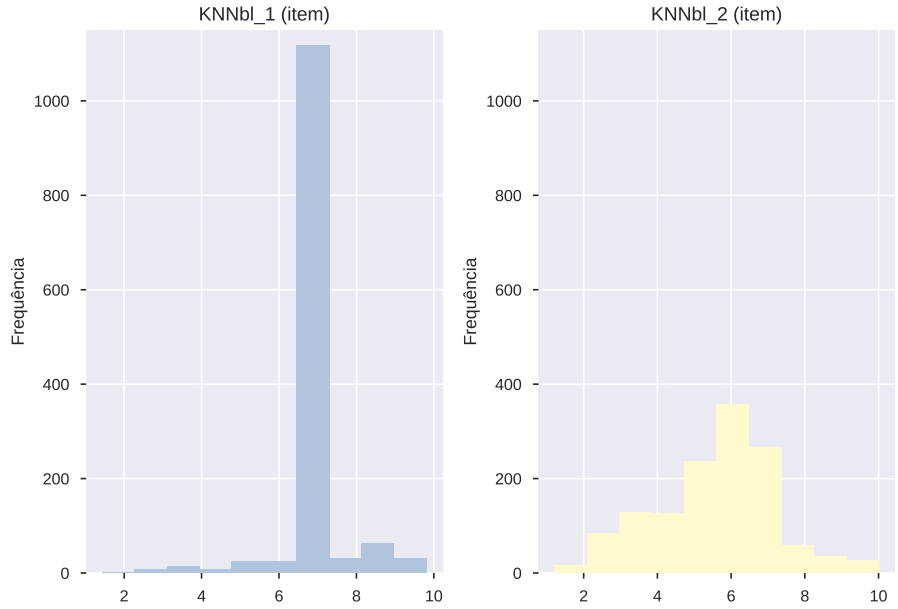
Tabela 6.3: Variação de parâmetros do algoritmo K-NN Baseline.

Parâmetro	Valores Testados	Descrição
k	[5, 10, 20, 40, 100]	Número (máximo) de vizinhos a serem considerados para agregação
min_k	[1, 5, 10]	Número mínimo de vizinhos a considerar para agregação
sim_options	name: [pearson_baseline] min_suport: [1, 5] method: [als, sgd]	Opções para a medida de similaridade
bsl_options	learning_rate': [0.00005, 0.02, 0.005, 0.1] n_epochs: [1,10,20,50] reg':[0.02,0.1,0.05]	Opções para o cálculo das estimativas da linha de base.

racterísticas semelhantes, ou seja, com o uso da métrica tradicional as predições ocorrem em maior número para notas mais elevadas. Enquanto que com a utilização a métrica proposta o maior número de predições ocorre para avaliações de menor valor. Em todos os casos, as previsões divergem entre as abordagens com a aplicação de cada métrica.

A Tabela 6.5 apresenta os resultados obtidos durante uma execução do processo de avaliação das métricas. Em R1 tem-se os resultados alcançados utilizando a avaliação da LBSN e em R2 estimativa gerada pela abordagem proposta. Nessa, são apresentados os resultados em termos do erro médio absoluto extraído. Dessa forma os resultados representam as respectivas médias e desvios padrão utilizando 10 *folds* para validação

Figura 6.6: Histograma de previsões para cada valor de predição utilizando K-NN Baseline baseado em item.



Fonte: Acervo pessoal.

cruzada dos algoritmos. De modo análogo, a Tabela 6.4 aponta resultados das respectivas médias e desvios padrão com 5 *folds* utilizados na validação cruzada de cada algoritmo utilizado. A disposição dos dados é análoga à utilizada na tabela anterior.

Tabela 6.4: Resultados de MAE obtidos utilizando validação cruzada com $k = 5$.

	R1[1]		R2[2]	
Algoritmo	Média	Desvio Padrão	Média	Desvio Padrão
SVD	0.277646	0.011833	0.324791	0.024981
SVD++	**0.201328**	**0.008558**	**0.249814**	**0.023112**
KNN (user)	**1.470982**	**0.051205**	**1.202484**	**0.035128**
KNN (item)	1.467083	0.048589	1.189732	0.034308
KNN Baseline (user)	1.243926	0.044831	1.043078	0.031952
KNN Baseline (item)	1.238402	0.042931	1.030933	0.031492

De modo a realizar uma validação mais consistente sobre os resultados obtidos, foram realizadas 30 execuções, para fins de validação estatística, empregando validação cruzada com 10 *folds*. Assim, a partir da extração das medidas de erro médio absoluto alcançadas utilizando cada abordagem de *rating*, realizou-se o teste de Wilcoxon, uma vez que este pode ser utilizado em amostras de métricas de avaliação de modelos averiguando se a diferença de predição entre os modelos de aprendizado de máquina é significativa.

[1]R1 refere-se a execução dos algoritmos de recomendação utilizando a nota proporcionada pela LBSN.
[2]R2 refere-se a execução dos algoritmos de recomendação utilizando a nota gerada a partir dos indicadores de recomendação extraídos.

Tabela 6.5: Resultados de MAE obtidos utilizando validação cruzada com $k = 10$.

	R1[1]		R2[2]	
	Média	Desvio Padrão	Média	Desvio Padrão
SVD	0.224628	0.010352	0.259426	0.008512
SVD++	**0.169321**	**0.010577**	0.215160	0.008744
KNN (user)	1.495882	0.033332	**1.220767**	**0.028342**
KNN (item)	1.040791	0.033670	0.866297	0.041528
KNN Baseline (user)	1.259964	0.030242	0.103477	0.003226
KNN Baseline (item)	**1.335254**	**0.032025**	**0.103352**	**0.003473**

A hipótese nula utilizada no teste consiste na suposição de que ambas as amostras foram retiradas de uma população com a mesma distribuição e, portanto, os mesmos parâmetros populacionais. Dessa forma, tem-se que se, após o cálculo do teste de significância no conjunto de amostras e a hipótese nula for rejeitada, há evidências que sugerem que as amostras foram retiradas de diferentes populações e, por sua vez, a diferença entre as estimativas amostrais pode ser significativa.

O teste retorna um valor p (*p-value*) que pode ser considerado como a probabilidade de observar as duas amostras de dados, considerando a hipótese nula. O *p-value* pode ser interpretado no contexto de um nível de significância de alfa. Dessa forma, se p estiver abaixo do nível de significância, então o teste diz que há evidências suficientes para rejeitar a hipótese nula e que as amostras provavelmente foram extraídas de populações com distribuições diferentes.

Ao realizar o teste de Wilcoxon sobre as amostras de dados, empregando um valor de alfa de 0.05, ou seja, 95% de confiança, foram atingidos os resultados retratados na Tabela 6.6. Para todos os algoritmos testados, a hipótese nula foi rejeitada. Isso representa a não uniformidade na distribuição das amostras, de modo que é possível afirmar que há diferenças significativas entre os resultados obtidos com ambas as abordagens de *rating* utilizadas.

Tabela 6.6: Resultados do teste Wilcoxon aplicado à MAE no decorrer de 30 execuções utilizando validação cruzada com $k = 10$.

Algoritmo	*p-value*	Hipótese
SVD	1.7343976283205784^{-06}	Rejeitada
SVD++	1.7343976283205784^{-06}	Rejeitada
KNN (user)	1.7343976283205784^{-06}	Rejeitada
KNN (item)	1.7343976283205784^{-06}	Rejeitada
KNN Base Line (user)	1.7343976283205784^{-06}	Rejeitada
KNN Base Line (item)	1.7343976283205784^{-06}	Rejeitada

Fonte: Acervo Pessoal

Também de modo a realizar uma validação mais consistente sobre os resultados obtidos, foram realizadas 30 execuções empregando validação cruzada com 5 *folds*. Assim, a partir da extração das medidas de raiz do erro médio quadrático alcançadas utilizando cada abordagem de *rating*, realizou-se sobre elas o teste de Wilcoxon.

A Tabela 6.7 apresenta os resultados obtidos durante uma execução do processo de avaliação das métricas. Nessa são apresentados os resultados em termos do erro médio absoluto extraído. Em R1 tem-se os resultados alcançados utilizando a avaliação fornecida pelo Foursquare e em R2 estimativa gerada pela estimativa gerada pela abordagem proposta. Desta forma, os resultados representam as respectivas médias e desvios padrão utilizando 5 *folds* para validação cruzada dos algoritmos. De modo análogo, a Tabela 6.8 aponta resultados das respectivas médias e desvios padrão com 10 *folds* utilizados na validação cruzada de cada algoritmo utilizado. Os dados encontram-se dispostos como: em R1 tem-se os resultados alcançados utilizando o *rating* fornecido pela LBSN e em R2 estimativa gerada pela abordagem proposta.

Tabela 6.7: Resultados de RMSE obtidos utilizando validação cruzada com $k = 5$.

	R1[1]		R2[2]	
Algoritmo	Média	Desvio Padrão	Média	Desvio Padrão
SVD	0.394302	0.044840	0.506881	0.019583
SVD++	**0.333497**	**0.056059**	0.458025	0.018890
KNN (user)	**1.942027**	**0.035910**	**1.573954**	**0.028715**
KNN (item)	1.633836	0.042253	1.348860	0.046112
KNN Baseline (user)	1.724629	0.032998	**0.212631**	**0.016364**
KNN Baseline (item)	1.758964	0.039059	0.212692	0.016421

Fonte: Acervo Pessoal

Tabela 6.8: Resultados de RMSE obtidos utilizando validação cruzada com $k = 10$.

	R1[1]		R2[2]	
Algoritmo	Média	Desvio Padrão	Média	Desvio Padrão
SVD	**0.349781**	**0.034429**	0.465482	0.027394
SVD++	0.408688	0.055431	0.412578	0.032038
KNN (user)	**1.938904**	**0.059910**	**1.567626**	**0.047501**
KNN (item)	1.583225	0.071083	1.292741	0.052349
KNN Baseline (user)	1.670258	0.053814	**0.191285**	**0.039884**
KNN Baseline (item)	1.728220	0.056223	0.191461	0.039523

Fonte: Acervo Pessoal

Ao realizar o teste de Wilcoxon sobre as amostras de dados, empregando um valor de alfa de 0.05, ou seja, 95% de confiança, foram atingidos os resultados retratados na

Tabela 6.9. Assim como no caso anterior, para todos os algoritmos testados, a hipótese nula foi rejeitada. Da mesma forma, isso representa a não uniformidade na distribuição das amostras, de modo que é possível afirmar que há diferenças entre os resultados obtidos com ambas as abordagens de *rating* utilizadas.

Tabela 6.9: Resultados do teste Wilcoxon aplicado a RMSE no decorrer de 30 execuções utilizando validação cruzada com $k = 10$.

Algoritmo	*p-value*	Hipótese
SVD	1.7343976283205784^{-06}	Rejeitada
SVD++	1.7343976283205784^{-06}	Rejeitada
KNN (user)	1.7343976283205784^{-06}	Rejeitada
KNN (item)	1.7343976283205784^{-06}	Rejeitada
KNN Base Line (user)	1.7343976283205784^{-06}	Rejeitada
KNN Base Line (item)	1.7343976283205784^{-06}	Rejeitada

Fonte: Acervo Pessoal

Ao realizar a comparação das métricas de MAE obtidas para cada algoritmo ao longo das 30 execuções desempenhadas, foi possível perceber que o uso da métrica de *rating* tradicional apresentou melhor desempenho em 100% dos casos, ao fazer uso de de algoritmos baseados em fatoração de matrizes (SDV e SVD++). Já para os algoritmos que são diretamente derivados de uma abordagem básica de vizinhos mais próximos, o uso da métrica proposta se mostrou melhor em todos os casos de comparação. A Tabela 6.10 apresenta a porcentagem de vezes em que cada abordagem se mostrou melhor quando comparada a outra.

Tabela 6.10: Comparação MAE

Algoritmo	R1[1]	R2[2]
SVD	100.00%	0.00%
SVD++	100.00%	0.00%
KNN (user)	3.33%	96.67%
KNN (item)	3.33%	96.67%
KNN Baseline (user)	3.33%	96.67%
KNN Baseline (item)	3.33%	96.67%

Fonte: Acervo Pessoal

Da mesma forma, ao comparar as métricas de RMSE obtidas para cada algoritmo ao longo das 30 execuções desempenhadas, foi possível perceber que o uso da métrica de *rating* tradicional apresentou melhor desempenho em todos os casos, ao fazer uso de de algoritmos baseados em fatoração de matrizes (SDV e SVD++). Já para os algoritmos que são diretamente derivados de uma abordagem básica de vizinhos mais próximos, o uso da

métrica proposta se mostrou melhor em todos os casos de comparação. A Tabela 6.11 apresenta a porcentagem de vezes em que cada abordagem se mostrou melhor quando comparada a outra.

Tabela 6.11: Comparação RMSE

Algoritmo	R1[1]	R2[2]
SVD	100.00%	0.00%
SVD++	100.00%	0.00%
KNN (user)	0.00%	100.00%
KNN (item)	0.00%	100.00%
KNN Baseline (user)	0.00%	100.00%
KNN Baseline (item)	0.00%	100.00%

Fonte: Acervo Pessoal

Com a realização de tais experimentos, foram obtidos melhores resultados com o uso da métrica proposta aplicando o uso de *features* extraídas a partir de análises em uma rede social baseada em localização, através do uso de algoritmos baseados em vizinhança. Entretanto, em casos onde o uso da métrica de *rating* comumente utilizada se mostrou melhor (algoritmos baseados em fatoração de matrizes), a taxa de erro medida com o uso de indicadores se mantém baixa como é possível observar nas Tabelas 6.7 e 6.8. Esse resultado acentua a aplicabilidade do uso de tais indicadores no processo de recomendação de equipamentos urbanos.

7 CONCLUSÃO

Este trabalho propôs uma análise de *features* que podem ser extraídas a partir de redes sociais baseadas em localização aplicando-as como indicadores no processo de recomendação de recursos urbanos. Para esste fim, foram propostas formas de extrair e quantificar indicadores tais como a polaridade dos comentários de avaliação de cada estabelecimento, a confiabilidade das informações providas por usuários baseando-se na quantidade de contribuições fornecidas previamente e, ainda, a relevância de *reviews* baseada no fator temporal (i.e., data de publicação). De modo a avaliar a estimativa de avaliação gerada a partir do uso de tais indicadores, foram realizados experimentos com algoritmos clássicos de sistemas de recomendação, buscando comparar essa com o uso do *rating* tradicional.

As abordagens de análise do conjunto de dados provenientes de uma rede social baseada em localização (Foursquare) destinavam-se a extração de informações consideradas relevantes para a avaliação de pontos de interesse. Dessa forma, foram estabelecidas equações para determinar os valores atribuídos à medidas de polaridade, confiabilidade e ao decréscimo de influência temporal. Assim, os experimentos realizados possuíam como objetivo averiguar métricas de acurácia atingidas por algoritmos de recomendação baseados em filtragem colaborativa (baseadas em vizinhança e baseadas em modelo) perante o uso da métrica acolhida.

De modo a melhor estudar o impacto da aplicação do uso dos indicadores de recomendação extraídos do conjunto de dados, os algoritmos analisados foram executados utilizando também as avaliações habituais utilizadas em LBSNs. Essas empregam em alguns casos a média aritmética das avaliações explícitas recebidas ou, ainda, como no caso do Foursquare, exploram brevemente outros aspectos. Foram realizadas 30 execuções dos algoritmos de vizinhança (K-NN e K-NN Baseline), baseados em item e em usuário e, ainda, de algoritmos de fatoração de matrizes (SVD e SVD++).

Ao avaliar o erro médio obtido durante a aplicação dos algoritmos de recomendação baseados em fatoração de matrizes, tem-se que em 100% dos casos o uso da avaliação padrão utilizada apresenta um erro inferior ao do uso da estimativa proposta. Entretanto, ressalta-se que nesses casos o erro se mantém com uma média e desvio padrão baixos. Já com a aplicação de algoritmos baseados em vizinhança, em todas as experimentações, o uso da estimativa sugerida se mantém com melhores resultados quando comparada à avaliação tradicional. Entretanto, nesses casos, para ambas as abordagens o erro se mostra

superior a um.

Ao avaliar a raiz do erro médio quadrático alcançada durante a aplicação dos algoritmos SVD e SVD++, assim como no caso anterior a estimativa de avaliação padrão, ela apresenta-se com melhores resultados, ou seja, menor taxa de erro (RMSE). Todavia, o valor de erro encontrado se mantém igualmente com uma média e desvio padrão considerados baixos. Assim, como ao considerar a MAE, a aplicação da abordagem de estimativa proposta demonstra resultados melhores em comparação com o uso do *rating* padrão.

Os resultados alcançados com o uso de ambas as métricas foram comparados utilizando testes de Wilcoxon com 95% de confiança. Esses retrataram a não uniformidade na distribuição das amostras, evidenciando diferenças significativas entre os resultados obtidos com as abordagens de *rating* utilizadas. Dessa forma, conclui-se que a aplicabilidade da estimativa proposta se mostra viável, uma vez que para as abordagens mais comumente utilizadas em pesquisas análogas voltadas a recomendação de elementos urbanos, são aplicados principalmente algoritmos baseados em vizinhança. Ainda, o uso em algoritmos realizados por processamento com fatoração não pode ser declinável uma vez que demonstraram uma boa acurácia, sendo ainda passíveis de aprimoramentos.

Posto que foram alcançados bons resultados no processo de recomendação de recursos urbanos, e tendo ainda que o uso de ontologias se mostra de grande importância para o desenvolvimento de aplicações voltadas ao cenário de cidades inteligentes, optou-se por dispor a implementação da população automatizada da ontologia tida por base. Dessa forma, torna-se possível então o desenvolvimento de sistemas de recomendação fundamentados em tal tipo de representação do conhecimento. Para tanto, tem-se ainda a possibilidade de integração da ontologia à aplicação do uso de indicadores de recomendação, visando a projeção de um sistema mais robusto e acurado.

Como trabalhos futuros, sugere-se então fazer uso da ontologia populada integrada aos indicadores gerados para aplicação a processos de recomendação de recursos urbanos, integrando ainda outras fontes de informação, propiciando o uso de *Web Mining* de estrutura de links da Web para o estudo do relacionamento entre páginas da Web através de seus hiperlinks, ou seja, visando o desenvolvimento de técnicas para aproveitar o julgamento coletivo da qualidade de páginas Web implícito na estrutura de ligações, propondo assim um sistema tido como mais completo, tornando possível a realização de testes *online*.

REFERÊNCIAS

ADOMAVICIUS, G.; TUZHILIN, A. Context-aware recommender systems. In: **Recommender systems handbook**. [S.l.]: Springer, 2011. p. 217–253.

AGGARWAL, C. C.; ZHAI, C. **Mining Text Data**. Springer New York, 2012. ISBN 9781461432234. Available from Internet: <https://books.google.com.br/books?id=vFHOx8wfSU0C>.

AMATRIAIN, X. Mining large streams of user data for personalized recommendations. **SIGKDD Explor. Newsl.**, ACM, New York, NY, USA, v. 14, n. 2, p. 37–48, abr. 2013. ISSN 1931-0145. Available from Internet: <http://doi.acm.org/10.1145/2481244.2481250>.

AMATRIAIN, X.; PUJOL, J. M. Recommender systems: Introduction and challenges. In: ____. **Recommender Systems Handbook**. Boston, MA: Springer US, 2015. p. 227–262. ISBN 978-1-4899-7637-6. Available from Internet: <https://doi.org/10.1007/978-1-4899-7637-6_1>.

BAGHERIFARD, K. et al. Performance improvement for recommender systems using ontology. **Telematics and Informatics**, v. 34, n. 8, p. 1772 – 1792, 2017. ISSN 0736-5853. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0736585317303404>.

BAO, J. et al. Recommendations in location-based social networks: a survey. **GeoInformatica**, v. 19, n. 3, p. 525–565, Jul 2015. ISSN 1573-7624. Available from Internet: <https://doi.org/10.1007/s10707-014-0220-8>.

Bellini, P. et al. Km4city ontology building vs data harvesting and cleaning for smart-city services. **Journal of Visual Languages and Computing**, v. 25, n. 6, p. 827–839, 2014.

CHAKRABARTI, S. **Mining the Web: Discovering Knowledge from Hypertext Data**. Elsevier Science, 2002. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780080511726. Available from Internet: <https://books.google.com.br/books?id=EtXSW9owrYYC>.

DAVENPORT, T. H. Analytics 3.0. **Harvard Business Review**, HARVARD BUSINESS SCHOOL PUBLISHING CORPORATION 300 NORTH BEACON STREET . . . , v. 91, n. 12, p. 64–+, 2013.

DELL EMC. 2017. Disponível em: https://www.emc.com/leadership/digital-universe/index.htm. Acesso em: 15 de dezembro de 2018.

DENTLER, K. et al. Comparison of reasoners for large ontologies in the owl 2 el profile. **Semantic Web**, IOS Press, v. 2, n. 2, p. 71–87, 2011.

EVANS, J. R.; LINDNER, C. H. Business analytics: the next frontier for decision sciences. **Decision Line**, v. 43, n. 2, p. 4–6, 2012.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: An overview. In: FAYYAD, U. M. et al. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Menlo Park, CA, USA: American Association

for Artificial Intelligence, 1996. p. 1–34. ISBN 0-262-56097-6. Available from Internet: <http://dl.acm.org/citation.cfm?id=257938.257942>.

FELDMAN, R.; DAGAN, I. Knowledge discovery in textual databases (kdt). In: **KDD**. [S.l.: s.n.], 1995. v. 95, p. 112–117.

FRAGOSO, S.; AMARAL, A.; RECUERO, R. **METODOS DE PESQUISA PARA INTERNET**. SULINA, 2011. ISBN 9788520505946. Available from Internet: <https://books.google.com.br/books?id=utWluAAACAAJ>.

GAMA, J. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo Gen - LTC, 2011. ISBN 9788521618805. Available from Internet: <https://books.google.com.br/books?id=4DwelAEACAAJ>.

GAO, H. et al. Content-aware point of interest recommendation on location-based social networks. In: **AAAI**. [S.l.: s.n.], 2015. p. 1721–1727.

GEMMIS, M. de et al. Semantics-aware content-based recommender systems. In: ____. **Recommender Systems Handbook**. Boston, MA: Springer US, 2015. p. 119–159. ISBN 978-1-4899-7637-6. Available from Internet: <https://doi.org/10.1007/978-1-4899-7637-6_4>.

GOLDSCHMIDT, E. P. e. E. B. R. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações**. [S.l.]: Elsevier Brasil, 2015.

GRACE, L. K. J. et al. Analysis of web logs and web user in web mining. **CoRR**, abs/1101.5668, 2011. Available from Internet: <http://arxiv.org/abs/1101.5668>.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowl. Acquis.**, Academic Press Ltd., London, UK, v. 5, n. 2, p. 199–220, jun. 1993. ISSN 1042-8143. Available from Internet: <http://dx.doi.org/10.1006/knac.1993.1008>.

GRUS, J. **Data Science from Scratch: First Principles with Python**. 1. ed. [S.l.]: O'Reilly Media, 2015.

GUIZZARDI, G. **Ontological foundations for structural conceptual models**. Thesis (PhD) — University of Twente, 10 2005.

HU, X.; LIU, H. Text analytics in social media. In: AGGARWAL, C. C.; ZHAI, C. (Ed.). **The Oxford Handbook of Innovation**. Springer: Springer, 2012. chp. 12, p. 385–414.

IBGE, I. B. de Geografia e E. **Gramado RS – Panorama**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística - IBGE, 2019. Available from Internet: <https://cidades.ibge.gov.br/brasil/rs/gramado/panorama>.

IBGE, I. B. de Geografia e E. **Projeção da população do Brasil e das Unidades da Federação**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística - IBGE, 2019. Available from Internet: <https://www.ibge.gov.br/apps/populacao/projecao/>.

ISINKAYE, F.; FOLAJIMI, Y.; OJOKOH, B. Recommendation systems: Principles, methods and evaluation. **Egyptian Informatics Journal**, v. 16, n. 3, p. 261 – 273, 2015. ISSN 1110-8665. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1110866515000341>.

ISOTANI, S.; BITTENCOURT, I. **Dados Abertos Conectados: Em busca da Web do Conhecimento**. NOVATEC, 2015. ISBN 9788575224496. Available from Internet: <https://books.google.com.br/books?id=TC9jCgAAQBAJ>.

JANNACH, D. et al. **Recommender Systems**. Cambridge: Cambridge University Press, 2010. ISSN 1044-7318. ISBN 9780511763113. Available from Internet: <http://ebooks.cambridge.org/ref/id/CBO9780511763113>.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Ijcai**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

KOMNINOS, N. et al. Smart city ontologies: Improving the effectiveness of smart city applications. **Journal of Smart Cities**, Whioce Publishing Pte Ltd, v. 1, n. 1, p. 16, nov 2016. Available from Internet: <https://doi.org/10.18063/jsc.2015.01.001>.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: AGGARWAL, C. C.; ZHAI, C. (Ed.). **The Oxford Handbook of Innovation**. Springer: Springer, 2012. chp. 13, p. 415–463.

LOPS, P.; GEMMIS, M. de; SEMERARO, G. Content-based recommender systems: State of the art and trends. In: ____. **Recommender Systems Handbook**. Boston, MA: Springer US, 2011. p. 73–105. ISBN 978-0-387-85820-3. Available from Internet: <https://doi.org/10.1007/978-0-387-85820-3_3>.

MAP, O. S. **OSM Semantic Network**. 2018. Available from Internet: <https://wiki.openstreetmap.org/wiki/OSM_Semantic_Network>.

MARQUES, J. R. B. e E. Cidades sustentáveis-inteligentes. In: **Projeto de ciência para o Brasil**. [S.l.]: Rio de Janeiro: Academia Brasileira de Ciências, 2017. v. 1, chp. 8, p. 185 – 205.

MEHMOOD, A. et al. Protection of big data privacy. **IEEE Access**, v. 4, p. 1821–1834, 2016. ISSN 2169-3536.

NAM, T.; PARDO, T. A. Conceptualizing smart city with dimensions of technology, people, and institutions. In: **Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times**. New York, NY, USA: ACM, 2011. (dg.o '11), p. 282–291. ISBN 978-1-4503-0762-8. Available from Internet: <http://doi.acm.org/10.1145/2037556.2037602>.

NORVIG, P.; RUSSELL, S. **Inteligência artificial: Tradução da 3a Edição**. Elsevier Editora Ltda., 2014. ISBN 9788535251418. Available from Internet: <https://books.google.com.br/books?id=BsNeAwAAQBAJ>.

NOY, N. F.; MCGUINNESS, D. L. et al. **Ontology development 101: A guide to creating your first ontology**. [S.l.]: Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, Stanford, CA, 2001.

PÉREZ-MARTÍNEZ, P. A.; MARTÍNEZ-BALLESTÉ, A.; SOLANAS, A. Privacy in smart cities-a case study of smart public parking. In: **PECCS**. [S.l.: s.n.], 2013. p. 55–59.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to recommender systems handbook. In: RICCI, F. et al. (Ed.). **Recommender Systems Handbook**. Boston, MA: Springer US, 2011. p. 1–35. ISBN 978-0-387-85820-3. Available from Internet: <https://doi.org/10.1007/978-0-387-85820-3_1>.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Recommender systems: Introduction and challenges. In: ____. **Recommender Systems Handbook**. Boston, MA: Springer US, 2015. p. 1–34. ISBN 978-1-4899-7637-6. Available from Internet: <https://doi.org/10.1007/978-1-4899-7637-6_1>.

RODRIGUEZ, J. D.; PEREZ, A.; LOZANO, J. A. Sensitivity analysis of k-fold cross validation in prediction error estimation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 3, p. 569–575, March 2010. ISSN 0162-8828.

SANTANA, B. S.; OLIVEIRA, J. Palazzo M de; WIVES, L. K. Modelos e sistemas para cidades inteligentes. **Cadernos de Informática**, Universidade Federal do Rio Grande do Sul, v. 10, n. 1, p. 10–14, 2018.

SARWAR, B. et al. Item-based collaborative filtering recommendation algorithms. In: **Proceedings of the 10th International Conference on World Wide Web**. New York, NY, USA: ACM, 2001. (WWW '01), p. 285–295. ISBN 1-58113-348-0. Available from Internet: <http://doi.acm.org/10.1145/371920.372071>.

SHMUELI, G.; KOPPIUS, O. R. Predictive analytics in information systems research. **Mis Quarterly**, JSTOR, p. 553–572, 2011.

SILVA, M. J. et al. **Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis**. [S.l.], 2010. Doi: 10455/6694. Available from Internet: <http://hdl.handle.net/10455/6694>.

SOUZA, M.; VIEIRA, R. Sentiment analysis on twitter data for portuguese language. In: **Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer-Verlag, 2012. (PROPOR'12), p. 241–247. ISBN 978-3-642-28884-5. Available from Internet: <http://dx.doi.org/10.1007/978-3-642-28885-2_28>.

STROHBACH, M. et al. Towards a big data analytics framework for iot and smart city applications. In: ____. **Modeling and Processing for Next-Generation Big-Data Technologies: With Applications and Case Studies**. Cham: Springer International Publishing, 2015. p. 257–282. ISBN 978-3-319-09177-8. Available from Internet: <https://doi.org/10.1007/978-3-319-09177-8_11>.

THAKURIAH, P. V.; TILAHUN, N. Y.; ZELLNER, M. Big data and urban informatics: Innovations and challenges to urban planning and knowledge discovery. In: ____. **Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics**. Cham: Springer International Publishing, 2017. p. 11–45. ISBN 978-3-319-40902-3. Available from Internet: <https://doi.org/10.1007/978-3-319-40902-3_2>.

TUZHILIN, A.; ADOMAVICIUS, G. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. **IEEE Transactions on Knowledge & Data Engineering**, v. 17, p. 734–749, 06 2005. ISSN 1041-4347. Available from Internet: <doi.ieeecomputersociety.org/10.1109/TKDE.2005.99>.

WAGA, K.; TABARCEA, A.; FRANTI, P. Recommendation of points of interest from user generated data collection. In: **2012 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2012)(COLLABORATECOM)**. [s.n.], 2013. v. 00, p. 550–555. Available from Internet: <doi.ieeecomputersociety.org/10.4108/icst.collaboratecom.2012.250451>.

WEISS, M. C.; BERNARDES, R. C.; CONSONI, F. L. Cidades inteligentes: casos e perspectivas para as cidades brasileiras. **Revista Tecnológica da Fatec Americana**, v. 5, n. 1, p. 01–13, 2017.

WIMALASURIYA, D. C.; DOU, D. Ontology-based information extraction: An introduction and a survey of current approaches. **Journal of Information Science**, v. 36, n. 3, p. 306–323, 2010. Available from Internet: <https://doi.org/10.1177/0165551509360123>.

YANG, D. et al. A sentiment-enhanced personalized location recommendation system. In: **Proceedings of the 24th ACM Conference on Hypertext and Social Media**. New York, NY, USA: ACM, 2013. (HT '13), p. 119–128. ISBN 978-1-4503-1967-6. Available from Internet: <http://doi.acm.org/10.1145/2481492.2481505>.

YANG, X. et al. A survey of collaborative filtering based social recommender systems. **Computer Communications**, Elsevier, v. 41, p. 1–10, 2014.

YING, J. J.-C. et al. Urban point-of-interest recommendation by mining user check-in behaviors. In: **Proceedings of the ACM SIGKDD International Workshop on Urban Computing**. New York, NY, USA: ACM, 2012. (UrbComp '12), p. 63–70. ISBN 978-1-4503-1542-5. Available from Internet: <http://doi.acm.org/10.1145/2346496.2346507>.

ZHANG, Y. Grorec: A group-centric intelligent recommender system integrating social, mobile and big data technologies. **IEEE Transactions on Services Computing**, v. 9, n. 5, p. 786–795, Sep. 2016. ISSN 1939-1374.

ZHENG, Y.; ZHOU, X. **Computing with spatial trajectories**. [S.l.]: Springer Science & Business Media, 2011.