UNIVERSIDADE ABERTA

Aplicação de Técnicas de *Data Mining* em Redes Sociais:

Estudo dos Próximos Locais a Visitar na Rede *Foursquare*

Eudália da Conceição Paraneta Bilro

Mestrado em Tecnologias e Sistemas Informáticos *Web*

Fevereiro de 2016

UNIVERSIDADE ABERTA

Aplicação de Técnicas de *Data Mining* em Redes Sociais:

Estudo dos Próximos Locais a Visitar na Rede *Foursquare*

Eudália da Conceição Paraneta Bilro

Mestrado em Tecnologias e Sistemas Informáticos *Web*

Trabalho de Projeto Orientado pelo

Professor Doutor Luís Manuel Pereira Sales Cavique Santos

Fevereiro de 2016

Resumo

A comunicação nas redes sociais surgiu da necessidade que o ser humano tem em partilhar assuntos, ideias, preferências comuns criando assim laços assentes em afinidades.

A constante presença dos utilizadores nas redes sociais expressando as suas opiniões sobre produtos, marcas, pessoas, gostos, ou costumes tem vindo a desencadear um grande interesse por parte de empresas e pessoas em analisar essas informações.

Numa sociedade que diariamente é capaz de produzir dados em massa, é cada vez mais necessária a criação de ferramentas para a sua análise e interpretação de forma a disponibilizar todo um conjunto de informações úteis para a tomada de decisões.

Neste contexto, este trabalho descreve o processo de aplicação de técnicas de Data Mining em dados extraídos da rede social *Foursquare* de forma a obter informações relevantes que auxiliem na identificação de padrão de comportamentos.

Através da descoberta de padrões sequenciais, este estudo irá permitir a visualização dos dados organizados numa poli-árvore com o objetivo de estudar os próximos locais a visitar na rede *Foursquare*.

Palavras-chave: Redes Sociais, Extração Dados, Padrões Sequenciais, *Foursquare*

Abstract

Communication in social networks has arisen from the need that the human being has to share subjects, ideas or common preferences and by this way creating bonds.

The constant presence of the users of social networks expressing their opinions about products, brands, people, tastes or habits has developed a great interest by companies and people to analyze this information.

In a society that is capable of, on a daily basis, producing mass information, it is necessary to create tools for the analysis and interpretation of such data as provide useful information to support the process of decision making.

In this context, this work describes the process of application of data and graph mining techniques in data extracted from the social network *Foursquare*, in order to obtain relevant information that support the identification of behavior patterns.

Through the discovery of sequential patterns, this study will allow to have a visualization of the data arranged in a “poly-tree” with the purpose of forecasting the next locations to be visited on the *Foursquare* network.

Keywords: Social networks, Data Mining, Sequential Patterns, *Foursquare*

Agradecimentos

Realizar um trabalho como este implica empenho, concentração, rigor e o acompanhamento e estímulo das pessoas que estão mais próximas. Na conclusão desta etapa da minha formação académica, não posso deixar de agradecer:

Ao meu orientador, o Professor Doutor Luís Cavique, pela sua experiência, sugestões, pela sua visão do todo, paciência nos momentos de indefinição e, acima de tudo, pela sua disponibilidade em ajudar;

À SAS Portugal, pela autorização de utilização das suas licenças que serviram de base à elaboração deste estudo.

Aos meus pais, que me ensinaram a vontade e a insatisfação pelo adquirido e me incentivaram sempre a procurar algo mais;

Ao meu companheiro Rúben de Carvalho, que me tem apoiado incondicionalmente nas minhas decisões;

Ao Guilherme e à Anna, por serem meus desafiadores, dando muito mais sentido à vida;

Ao meu sobrinho Ivo Figueira, cuja disponibilidade e extraordinário apoio possibilitaram este projeto;

Aos meus amigos, a todos, o meu MUITO OBRIGADA!

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

1.1. Introdução

Uma rede social pode ser definida como um conjunto de pessoas ou grupos de pessoas, ligados entre si por relações de vários tipos e que partilham valores e objetivos comuns.

O grande impulsionador da Análise de Redes Sociais foi Moreno [1934], na década de 30, que partiu dos pressupostos da sociometria com o uso do seu instrumento de recolha de informação, o sociograma, sendo o teste sociométrico a primeira estratégia utilizada para conseguir analisar a estrutura de um grupo.

O matemático Anatole Rapoport [1957] começou nos anos 50 a traçar uma visão da sociedade como uma rede de pessoas, cada uma delas com vínculos aleatórios com outras pessoas. Esses vínculos poderiam ser curtos, isto é, ligar pessoas em comunidades próximas, ou longos, que estabeleceriam ligações com pessoas de comunidades mais distantes.

Contudo, foi só com os trabalhos do psicólogo Stanley Milgram [1967], nos anos 60 que a ideia de que a distância, medida em número de ligações de conhecimento direto, ou graus de separação, entre dois elementos típicos de uma rede de ligações sociais é de facto bastante pequena, mesmo em redes com muitos elementos, como a sociedade americana dos anos sessenta.

Atualmente, e em virtude das novas tecnologias, o acesso às redes sociais assume uma forte tendência, principalmente no que respeita às redes baseadas em serviços de localização. Geralmente, trata-se de aplicações utilizadas em *smartphones* com acesso à Internet e ao GPS, em que os utilizadores podem partilhar referências a locais, num determinado sistema. As suas funcionalidades passam por permitir aos seus utilizadores a partilha do lugar onde se encontram, ligar-se a pessoas que estão no mesmo local, visualizar comentários de utilizadores com os mesmos interesses e pesquisar locais através de informações relevantes.

Tendo em vista este contexto, este trabalho irá focar-se na maior rede social baseada em geolocalizacão – o *Foursquare* [2014], uma plataforma lançada em 2009, com mais de 10 milhões de utilizadores e perto de 1 milhão de *check-ins* por dia.

1.2. Objetivos

Neste trabalho é apresentado um estudo do comportamento de padrões sequenciais através dos *check-ins* de uma aplicação, que irá permitir definir e prever hábitos de localização futuros dos seus utilizadores. O seu objetivo consiste em analisar um conjunto de ferramentas utilizadas na extração, transformação e processamento de dados de redes sociais, com o intuito de definir e prever hábitos de localização de utilizadores, usando como fonte de análise uma plataforma de localização móvel - o *Foursquare* [2014].

Para tal foi conduzido um estudo abrangente sobre as principais funcionalidades do *Foursquare*, passando por uma descrição estatística dos utilizadores, concluindo-se com o estudo e aplicação do Algoritmo Ramex [Cavique 2007, Cavique e Coelho 2008, Cavique 2015], de forma a estudar o próximo local a ser visitado.

A motivação para este estudo surge pelo facto de contemplar uma área bastante recente, aliada à capacidade que estas redes apresentam de coletarem dados que podem vir a constituir valiosas fontes de informação na área de estudo do comportamento humano. Os dados recolhidos, podem ser analisados segundo a frequência de determinados eventos, ou conjuntos de eventos, relacionados com particularidades temporais. Observa-se, também, que estas tipologias de análises podem ser muito úteis para predizer o comportamento futuro de um processo monitorizado.

Embora existam já vários estudos neste âmbito, os quais sugerem orientações para as análises sobre as interações dos utilizadores nas redes social baseadas em geolocalizacão, nenhum deles é focado no uso de algoritmos que permitam uma visão global dos dados e que possibilitem a visualização dos dados organizados numa poli-árvore.

As reflexões deste estudo visam explicar, prever e controlar o comportamento do indivíduo nas redes sociais, de forma a explicar factos futuros.

1.3. Estrutura do Trabalho de Projeto

Este trabalho encontra-se dividido em sete capítulos. No presente capítulo, procede-se à identificação do estudo realizado e são formulados os seus objetivos.

No segundo capítulo é feita uma revisão da literatura existente sobre a determinação do próximo local.

Já no terceiro capítulo é apresentada uma breve introdução sobre a rede *Foursquare*, definindo conceitos importantes para um melhor entendimento do tema apresentado.

O processo de Transformação dos Dados do *Dataset Foursquare* é abordada no capítulo quatro.

O quinto capítulo apresenta uma análise à rede *Foursquare* de forma a obter uma caracterização sobre a arquitetura da rede, identificando *hubs* e possíveis ligações entre locais.

No sexto capítulo são apresentadas a análise e discussão dos resultados, assim como a interpretação e diagnóstico acerca da temática.

Por fim, no sétimo e último capítulo são expostas as conclusões relativamente ao tema e objetivos definidos no início do trabalho.

2. Trabalhos Relacionados

2.1. Análise da Rede Social Foursquare

A Análise de Redes Sociais constitui hoje um tema de investigação em grande desenvolvimento. Este facto deve-se sobretudo à sua atualidade, ao ritmo a que se desenvolve, e às múltiplas questões que esta recente temática pode levantar.

Tendo em vista esse contexto, o presente capítulo apresenta um levantamento bibliográfico sobre redes sociais baseadas em localização, dando ênfase a pesquisas realizadas acerca do Foursquare, uma das principais redes desta natureza.

2.1.1. Literatura Relacionada com a Rede Foursquare

Existem vários trabalhos que caracterizam diferentes aspetos da rede social Foursquare. Um dos primeiros estudos nesta área conseguiu demonstrar a existência de ligações de amizade entre curtas distâncias estabelecendo relações entre propriedades sociais e geográficas nas redes sociais [Scellato et al., 2010].

Em [Noulas et al., 2011] foi feita uma analise à dinâmica dos *check-ins*, os quais mostram padrões dispostos ao longo do tempo e do espaço, nomeadamente os padrões espácio-temporais e a mobilidade dos utilizadores nos espaços urbanos. Os mesmos autores [Mascolo et al., 2011] mas num outro estudo, utilizaram técnicas de agrupamento para identificar comunidades e caracterizar os vários tipos de atividades que acontecem em cada um desses grupos.

O estudo exposto em [Vasconcelos et al., 2012] apresenta uma caracterização do modo como os utilizadores interagem entre si por meio de opiniões, através da recolha dos seus percursos no *Foursquare*.

Um estudo um pouco diferente, realizado em [Sadilek & Krumm, 2012], propõe um modelo de aprendizagem baseado nos lugares visitados pelos utilizadores com recurso a um método de validação cruzada. De forma semelhante, [Mascolo et al., 2011] apresenta um estudo sobre os lugares visitados pelos utilizadores, mas tendo em conta a variação de tempo e espaço.

Finalmente, em [Pietro & Cohn 2013] adota-se uma visão centrada no utilizador, em que foram estudados padrões de mobilidade combinando a informação temporal e

espacial, de forma a demonstrar: 1) como agrupar utilizadores baseados no seu comportamento; 2) como melhorar a previsão de locais futuros com recurso à análise de consecutivos *check-ins* do mesmo utilizador.

2.1.2. DataSet Foursquare

O processo de descoberta de padrões foi aplicado em uma amostra de dados reais referentes aos check-ins da aplicação *Foursquare*.

Para a realização deste estudo recorreu-se a dados primários disponibilizados em http://www-users.cs.umn.edu/~baojie/Research.htm que serviram de base para um estudo que apresenta um sistema de recomendação baseado nas preferências pessoais de cada utilizador [Bao, Zheng & Mohamed, 2012]. Os dados foram cedidos por um dos autores, juntamente com um *email* que autoriza a utilização do *dataset* para o presente estudo.

No capitulo seguinte são descritos as funcionalidades da ferramenta Foursquare bem como o modelo relacional e a estatística descritiva do *dataset*.

2.2. Descoberta de Padrões Sequenciais

Uma das atividades de *Data Mining* é a descoberta de padrões sequenciais.

A criação de algoritmos eficientes para a descoberta do conhecimento tem crescido de forma significativa devido à existência de grandes volumes de dados com dependência temporal e decorre da necessidade iminente de extrair conhecimento dessa mesma informação, uma vez que o ser humano não é capaz de interpretar uma grande quantidade de dados [Fayyad et al., 1996].

O problema da descoberta de padrões sequenciais foi inicialmente discutido em [Agrawal & Srikant, 1995]. Considerando uma base de dados de sequências, onde cada sequência é uma lista de transações ordenadas pelo tempo e cada transação uma lista de itens, o problema relacionado com a descoberta de padrões sequenciais consiste em encontrar subsequências frequentes que satisfaçam os critérios mínimos especificados pelo utilizador.

Uma sequência composta por uma série de valores reais, medidos em vários pontos do tempo, é comummente chamada série temporal. Uma característica que distingue os dados de uma série temporal de outros tipos de dados é que, em geral, os valores de uma série, em diferentes instantes do tempo, são correlacionados [Lin, Orgun & Williams, 2002].

Sabendo-se que uma série temporal é uma descrição do passado, um procedimento lógico para realizar previsões será a utilização desses mesmos dados históricos. Se os dados passados são indicativos do que se pode esperar do futuro, pode-se então implementar um modelo matemático que represente o processo; o modelo gerado pode, assim, ser usado para realizar previsões. Com efeito, os modelos de previsão que se apoiam principalmente na evolução dos dados históricos tendem a assumir que o comportamento futuro não se desviará significativamente da tendência verificada [Zarur, 2005].

Segundo o estudo apresentado em [Cavique & Coelho, 2008], a prospeção de padrões temporais pode ser dividida em quatro diferentes abordagens: os padrões periódicos, a descoberta de sequências, os episódios frequentes e os modelos das cadeias de Markov.

[Knuth, Morris & Pratt, 1977] propuseram um algoritmo para o reconhecimento de padrões em texto que, no pior caso, efetua O(m+n) comparações, sem realizar retrocessos no texto. Além disso, cada caractere do texto é comparado, no máximo, O(log m) vezes e, supondo que o texto é armazenado externamente, este algoritmo utiliza apenas O(m) posições de memória.

Em [Mannila, Toivonen & Verkamo, 1997], coloca-se o problema de encontrar episódios frequentes numa longa sequência de eventos - um episódio define-se como um conjunto de eventos ocorrendo numa ordem parcialmente definida, num intervalo de tempo. Para serem considerados interessantes, os eventos de um determinado episódio devem ocorrer suficientemente próximos no tempo. O utilizador define o grau de proximidade, fornecendo a largura da janela de tempo e a periodicidade com que o mesmo deve ocorrer, para ser considerado frequente.

A abordagem principal à descoberta de padrões entre dados com dependências temporais consiste na adaptação do algoritmo tradicional de descoberta de regras de associação (*Apriori*) ao caso particular deste tipo de dados. Alguns exemplos são os algoritmos AprioriAll [Srikant & Agrawal, 1995] e GSP (*Generalized Sequential Pattern*).

Este capítulo irá abordar os modelos e algoritmos mais relevantes bem como as suas contribuições para a técnica da descoberta de padrões sequenciais. A mesma que, é apresentada como base na abordagem ao algoritmo Ramex [Cavique 2007, Cavique e Coelho 2008, Cavique 2015] e cuja sua utilização permitirá detetar padrões pelo conceito de poli-árvores.

2.2.1. Apriori

O *Apriori* é um algoritmo que encontra regras de associação, sem se preocupar com a ordem temporal dos itens destas regras. Pra encontrar as regras, é realizado um procedimento iterativo, onde cada iteração executa duas ações: geração de candidatos possivelmente frequentes e definição dos padrões frequentes. Para avaliar se uma regra deve ou não ser considerada, são utilizadas duas medidas: suporte e confiança. São usados como filtros, para diminuir o número de regras geradas, gerando apenas regras de melhor qualidade.

O suporte é a probabilidade do antecedente da regra estar presente na base em relação a quantidade total de registros na base, ou seja,

$$S = \frac{X_q}{T}$$

Onde X_q é a quantidade de vezes que o item X aparece na base, e T é o total de itens diferentes da base. Já a confiança, é a probabilidade de o consequente e antecedente estarem presentes juntos relativo aos registros em que aparece o antecedente. Ou seja:

$$S = \frac{XY_q}{X_q}$$

Onde XY_q é a quantidade de sequências em que os itens X e Y aparecem na base, e Xq é a quantidade de vezes que o item X aparece. Tendo estas duas medidas definidas,

o algoritmo começa por verificar os itens que atendem ao suporte mínimo. Ou seja, para um suporte mínimo de 0,5, os itens que aparecerem em pelo menos em metade das sequências na base de dados, serão considerados candidatos frequentes e passarão para a próxima fase.

Sejam I e J dois itemsets tais que I está contido em J. Se J é frequente então I também é frequente. Sendo assim, os itens que são podados em uma iteração, impedem que uma gama de candidatos certamente infrequentes sejam criados, o que determina que, para que um itemset seja frequente, é necessário que todos os itemsets contidos nele também sejam. Se existir pelo menos um que não satisfaça ao suporte mínimo, então sabemos de antemão que qualquer candidato a partir dele não precisa ter suporte calculado, pois certamente não será frequente.

Partindo de uma base de dados onde estão listadas algumas sequências de itens em diferentes transações, cada iteração possui como candidatos conjuntos de itens (chamados itemsets) misturados de acordo com os considerados frequentes na transação anterior. Aqueles que não atenderem ao suporte mínimo são podados e não geram candidatos na próxima iteração. Os resultados que possuem tamanho maior, tornam-se as regras de associação do algoritmo.

Cada regra então passa por uma verificação de confiança. Aquelas que atenderem à confiança mínima, são as regras resultantes do processo.

Algoritmo 1 - Algoritmo *Apriori*

C_k: itemset candidato tamanho k
L_k : itemset frequente de tamanho k

L_1 = {itens frequentes de tamanho 1};
for (k = 1; L_k !=∅; k++) **do begin**
  C_{k+1} = gerar itemset candidatos L_k;
  for each transacao t na basedados **do**
    incrementar num de candidados em C_{k+1}g que estão contidos em t
  L_{k+1} = candidatos em C_{k+1} com min_supporte
  end
return $\cup_k L_k$;

2.2.2. AprioriAll

O algoritmo *AprioriAll* aproveita as premissas desenvolvidas no *Apriori*, com a diferença de que neste, a ordem com que os itens aparecem é de total importância.

O objetivo é o de encontrar os itens que costumam aparecer na base após o aparecimento de outros itens. A base de dados de um algoritmo deste tipo necessita - além da listagem dos itens - da correlação entre o *item* e a data determinante desta transação.

O processo é dividido em duas partes: A primeira está relacionada com a preparação e transformação da base de dados e segunda é a própria execução do algoritmo *AprioriAll*.

Na primeira fase, a base de dados original sofre as primeiras transformações de forma a criar transações agrupadas em uma só sequência, na qual os *itemsets* aparecem seguindo uma ordem cronológica, começando do mais antigo, e terminando com o mais recente. Feito isso, são listados os *itemsets* separadamente para o cálculo do suporte. Os que não atingirem suporte mínimo são podados e não passam para a fase de mapeamento, onde todos os *itemsets*, únicos ou não, são mapeados em valores numéricos utilizados na próxima fase.

Na fase seguinte, a base de dados é transformada de acordo com os *itemsets* que sofreram mapeamento. Neste passo, aqueles considerados infrequentes na etapa anterior serão retirados da base. Com a nova base, é necessário refazer o cálculo do suporte, visto que os *itemsets* agora sofreram modificações. Novas sequências serão consideradas frequentes, e depois mapeadas novamente para o seu valor original, retornando os padrões frequentes da base.

2.2.3. GSP (*Generalized Sequential Pattern*)

O algoritmo GSP difere do *AprioriAll* principalmente nas etapas de criação de candidatos e poda dos candidatos. Nesta última, são podados muito mais candidatos por iteração, devido a uma otimização na construção de seus pré-candidatos.

No algoritmo *AprioriAll*, em cada iteração k, os conjuntos Lk e Ck (*Itemsets* frequentes e *itemsets* candidatos) são constituídos de sequências de k *itemsets*.

No algoritmo GSP, em cada iteração k os conjuntos Lk e Ck (*Itemsets* frequentes e *itemsets* candidatos) são constituídos de sequências de k itens.

Ou seja, os *itemsets* frequentes <A> e <B> dão origem, no *AprioriAll*, ao candidato <A, B>. Já no algoritmo GSP, os mesmos dão origem a dois candidatos:

<A, B> e <A, B> ou seja, ao invés de darem origem a um candidato que possui dois *itemsets* (conjunto de itens), dá origem a dois candidatos que possuem dois itens, estejam eles em *itemsets* distintos ou não.

2.2.4. Cadeias de Markov

A cadeia de Markov pode ser vista como um sistema de estados e transições. É um caso particular de processo estocástico, com tempo discreto, que segue a propriedade de memória *markoviana*, onde os estados anteriores são irrelevantes para a predição dos estados seguintes, desde que o estado atual seja conhecido.

Uma cadeia de Markov é uma sequência X1, X2, X3,.. de variáveis aleatórias. O conjunto de valores que elas podem assumir é designado como espaço de estados, em que Xn denota o estado do processo no tempo n. Se a distribuição de probabilidade condicional de X^{n+1} nos estados passados é apenas uma função de X_n, então:

$Pr\ (X_{n+1} = x|X_0,X_1,X_2,\ldots.X_n) = Pr\ (X_{n+1} = x|\ X_n)$,onde X é algum estado do processo.

Os tipos de cadeia de Markov, finitas e discretas, podem também ser descritas por meio de um grafo dirigido (orientado). Aqui, cada aresta é rotulada com as probabilidades de transição de um estado para outro, sendo estes estados representados como os nós conectados pelas arestas. Estas probabilidades de transição são normalmente agrupadas numa matriz de transição, onde o (i, j)-ésimo elemento é igual a

$P_{ij} = Pr\ (X_{n+1} = j \mid X_n = i)$.

Para um espaço de estados discretos, as integrações na probabilidade de transição de *k* passos são somatórios e podem ser calculados como a *k*-ésima potência da matriz de transição. Isto é, se **P** é a matriz de transição para um passo, então $\mathbf{P}^k$ é a matriz de transição para a transição de *k* passos [Wikipedia, 2014].

2.2.5. Algoritmo Ramex

O nome do algoritmo *Ramex* provém do Latim*,* que significa “ramo”, em português. Esta abordagem apresenta a informação em rede, pois todos os itens são levados em consideração, e permite uma visão global dos dados. O seu objetivo é criar uma sequência de árvores com várias ramificações, tantas quantas as necessárias, de forma a garantir que todos os ramos serão visitados a partir do seu vértice-raiz.

2.2.5.1. Conceitos Fundamentais

Em muitos problemas que nos surgem no dia-a-dia, a forma mais simples de os descrever é representá-los em forma de grafo, uma vez que um grafo oferece uma representação visual que trará vantagens na construção de um modelo matemático com vista à resolução do problema.

A teoria dos grafos pode ser considerada um ramo recente da matemática. Nas estruturas designadas como grafos - G (V,A) -, V é um conjunto não vazio de objetos, denominados vértices, e A é um conjunto de pares não ordenados de V, chamado arestas. Um grafo não direcionado (ou simplesmente grafo) é dado por: um conjunto V de vértices; um conjunto E de arestas;

Na computação, um grafo finito direcionado ou não-direcionado (com n vértices) é geralmente representado pela sua matriz de adjacência: uma matriz n-por-n, cujo valor na linha i e coluna j fornece o número de arestas do i-ésimo ao j-ésimo vértices.

Uma árvore é um grafo orientado com um e um só caminho simples entre quaisquer dois vértices. Um subgrafo que seja uma árvore e contenha todos os vértices do grafo é designado por árvore abrangente.

Dado um grafo não orientado conectado, uma árvore de extensão deste grafo é um subgrafo, que é uma árvore que conecta todos os vértices. Um único grafo pode ter diferentes árvores de extensão. Existe ainda a possibilidade de associar um peso a cada aresta, número que representa a quão desfavorável ela é, e a possibilidade de atribuir um peso à árvore de extensão, calculado pela soma dos pesos das arestas que a compõem. Uma árvore de extensão mínima (também conhecida como árvore de

extensão de peso mínimo ou árvore geradora mínima) é, então, uma árvore de extensão com peso menor ou igual a cada uma das outras árvores de extensão possíveis.

O problema da Árvore Geradora Mínima consiste em encontrar uma árvore geradora de custo mínimo num grafo conexo através de arestas com custos para conectar os seus vértices. A árvore encontrada é aquela com o menor custo possível de entre todas as árvores geradoras possíveis para um determinado grafo.

O primeiro algoritmo a encontrar uma árvore de extensão mínima foi desenvolvido pelo cientista checo Otakar Borůvka, em 1926. Existem outros dois algoritmos habitualmente usados, o algoritmo de Prim e o algoritmo de Kruskal. Classificados como algoritmos heurísticos, ambos resolvem problemas de otimização, realizando a melhor escolha no momento e expandindo os nós que se encontram mais próximos do objetivo final.

No algoritmo de Krustal, o principal objetivo é o de selecionar a melhor aresta sem preocupações com as arestas selecionadas anteriormente. O resultado é uma proliferação de árvores que eventualmente se juntam para formar uma só. Baseado nesse prossuposto, o objetivo de Prim (Algoritmo 1) foi pensado para que a árvore cresça naturalmente até à obtenção da árvore geradora mínima. Assim, a próxima aresta selecionada seria sempre uma que se conecta à árvore que já existe. No início, o conjunto contém um vértice arbitrário. A cada passo, o algoritmo considera todas as arestas que tocam nesse conjunto e seleciona a aresta segura. Para a escolha de uma aresta segura deve-se observar o conjunto de arestas possíveis e selecionar aquelas que não formam ciclos com o subgrafo até então formado e cujo peso é o mínimo possível naquele momento. De seguida, o algoritmo acrescenta ao conjunto inicial o vértice ligado por essa aresta, que não estava no próprio conjunto. O processo continua até que o conjunto contenha todos os vértices da rede.

O fluxo máximo é o fluxo de maior valor possível. Numa rede G, os valores associados aos arcos desta rede representam as respetivas capacidades, isto é, a quantidade máxima de fluxo que pode ser enviada pelos arcos.

Em problemas de fluxo máximo, existem 2 nós especiais: nó de origem e nó de destino. Com a resolução do problema de fluxo máximo, pretende-se determinar a

quantidade máxima de unidades de fluxo que podem ser enviados de um nó de origem S para um nó terminal T.

Algoritmo 2 - Algoritmo de Prim

```
Escolha um vértice S para iniciar o subgrafo
Enquanto houver vértices que não estão no subgrafo
    Selecione uma aresta segura
    Insira a aresta segura e seu vértice no subgrafo
```

O algoritmo de Ford e Fulkerson é um procedimento iterativo baseado no problema do fluxo máximo. O processo inicia-se com um fluxo viável (igual a zero, quando não conhecido) de S para T e procura-se um caminho de aumento de fluxo. Se este caminho for encontrado, então enviam-se tantas unidades de fluxo quantas for possível por este caminho. Procura-se, então, novamente, um outro caminho de aumento de fluxo de S para T e assim sucessivamente, até que não haja nenhum outro caminho e, neste caso, o fluxo corrente é máximo.

Este algoritmo apresenta complexidade pseudopolinomial de O(n.m.U), para um grafo de n vértice e m arcos, onde U é o valor da maior capacidade de um arco do grafo.

2.2.5.2. Fases do Algoritmo Ramex

A abordagem do Algoritmo Ramex [Cavique 2007, Cavique e Coelho 2008, Cavique 2015] tem duas fases. Na primeira fase, dá-se a transformação da base de dados numa rede. Os dados são ordenados e, para cada linha, um novo atributo é criado: o próximo item. Em seguida é criada uma rede, i.e., um grafo com uma fonte (ou raiz) e um nó final chamado sumidouro. Esta rede, em que os ciclos são permitidos, condensa a informação da base de dados, incorporando todas as sequências possíveis. Na rede construída, cada nó corresponde a um item e a transição representa uma sequência de um item para o próximo item. O peso de cada arco corresponde ao número de vezes que um item antecede um próximo item. A transformação da base de dados numa rede é idêntica à abordagem das Cadeias de Markov.

A segunda fase, com base na rede cíclica, é onde é criada a árvore das sequências mais frequentes.

O Algoritmo 3 descreve-se em duas fases: a fase de transformação da base de dados numa rede cíclica e a procura da árvore das sequências frequentes.

Algoritmo 3 - Algoritmo Ramex

Input: Uma base de dados
Output: árvores (ou poli-árvore) de itens sequenciais;
1) Fase de Transformação em Rede:
- Ordenação dos dados;
- Criação de um novo atributo próximo item;
- Construção da rede com base nas sequências;

2) Fase de Procura:
- Procura na rede cíclica a árvore de itens mais frequente, utilizando a heurística *Back-and-Forward*.

O algoritmo *Ramex*, tal como o modelo das cadeias de Markov, não necessita de parâmetros e apresenta uma visão global, onde todos os itens são levados em conta, condensando a informação numa rede cíclica. Em cada transação são usadas frequências absolutas, reportando o número de vezes que um item antecede um próximo item. Após a construção da rede, será efetuada uma pesquisa para procurar uma árvore que represente a sequência provável dos itens. O *output* final será uma poli-árvore que represente a sequência encontrada.

Na primeira abordagem ao algoritmo *Ramex* [Cavique 2007], a Árvore de Sequências resultava da aplicação do algoritmo de Fulkerson [1974] à rede, na qual os dados eram condensados em uma base de dados.

Em [Cavique & Coelho, 2008], para a geração de poli-árvores, desenvolveu-se a heurística *Back-and-Forward*, também baseada no algoritmo de Prim, cuja complexidade temporal é representada por $O(N^2)$, em que N representa o número de vértices.

2.2.5.3. Heurística *Back-and-Forward*

A poli-árvore é um grafo orientado acíclico, com um arco entre cada par de nós no máximo. O grau interno dos vértices de uma árvore é zero (a raiz) ou um. Por sua vez, o grau interno dos vértices de uma poli-árvore pode ser maior do que um.

Os dois algoritmos, de [Edmonds, 1967] e [Fulkerson, 1974], geram árvores, com grau interno dos vértices de zero ou um. Para encontrar a poli-árvore ponderada de maior peso não existe nenhum algoritmo polinomial conhecido, daí o uso de heurísticas nesta fase como método simplificador do algoritmo. O algoritmo 3 demonstra o processo para a Heurística *Back-and-Forward.*

Algoritmo 4 - Heurística *Back-and-Forward*

```
Input: Rede G;
Output: Árvore S;
Iniciar S;
Para cada vértice em G
        Para cada arco em G
                x = arg_max (vértice ponderado à frente, não visitado em G e ligado a S )
                vértice ponderado atrás, não visitado em G e ligado a S)
        Fim-para;
        Atualizar S com x;
Fim-para.
```

A figura 2.1 mostra um exemplo de uma rede cíclica à qual será aplicada a heurística Back-and-Forward.

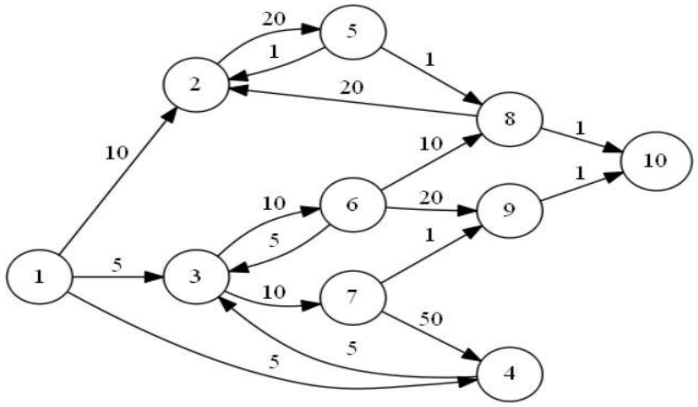

Figura 2.1 - Exemplo de rede cíclica

A árvore ponderada de maior peso (árvore que visita todos os nós) tem um comprimento total igual a 151 (50 + 10 + 20+ 10 + 20 + 20 + 10+1) e é representada na Figura 2.2.

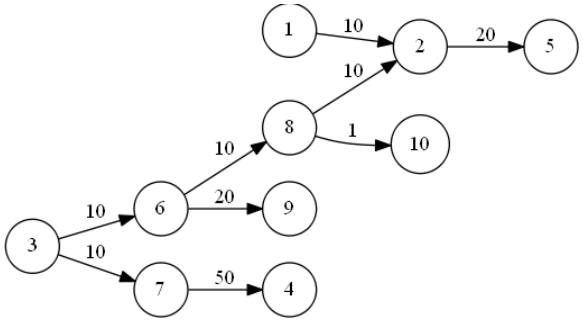

Figura 2.2 - Poli-árvore de maior peso

2.3. Ferramenta de Análise e Visualização

O processo de descoberta de conhecimento pode ser fortalecido pela incorporação de técnicas de visualização. As representações visuais manipuladas dinamicamente facilitam a exploração e interpretação dos dados pelo utilizador, possibilitando a perceção de padrões, tendências, relacionamentos e exceções embutidos nos mesmos.

Foi neste contexto que surgiu a necessidade de encontrar ferramentas que conseguissem uma representação gráfica legível da rede e da poli-árvore resultante.

Para obter medidas referentes à rede *Foursquare*, oferecendo uma caracterização sobre a sua arquitetura e ao mesmo tempo a sua representação, foi utilizado um S*oftware* de código aberto baseado em Java, o GEPHI [Gephi, 2015].

É uma das principais ferramentas de análise e visualizações de rede. Permite a visualização de sistemas complexos e tem a capacidade de processar redes de grande dimensão dependendo apenas da capacidade do computador. É gratuita, de código aberto e continuamente aperfeiçoado através da colaboração de milhares de utilizadores de todo o mundo.

Após testar várias soluções, o *Graphviz* demonstrou ser a ferramenta mais completa em termos de criação, análise e exploração da poli-árvore, muito devido sobretudo à simplicidade e agilidade do processo de visualização.

Graphviz [Graphviz, 2015] permite a visualização de grafos através de um ficheiro de texto escrito na linguagem Dot. É uma ferramenta focada para a visualização, pois é possível modificar completamente o *layout* do grafo, alterando as cores e imagens dos nós. Embora não seja possível interferir na disposição gráfica do diagrama gerado, o algoritmo utilizado é muito bom e consegue organizar os nós numa disposição aceitável.

2.4. Conclusão

Foram descritos neste capítulo os algoritmos e métodos mais relevantes de forma a enquadrar o tema principal deste estudo. Os pontos abordados permitiram compreender conceitos como o de descobrir, iterativamente, os padrões ou sequências frequentes existentes em bases de dados.

Dado um grafo não orientado e conectado, o Algoritmo de Prim permite encontrar a árvore geradora ("*spanning tree*") mínima ou máxima, originando um subgrafo que contém todos os vértices.

A Heurística *Back-and-Forward* utilizada no Algoritmo Ramex encontra a árvore ponderada de maior peso que visita todos os nós, aplicando a mesma técnica para grafos dirigidos. A complexidade temporal da Heuristica *Back-and-Forward* e para o Algoritmo de Prim é igual , e é representada por Θ(E + V. log V) onde E representa o numero de arestas e V o número de vértices [Cavique, 2015].

O Algoritmo Ramex tem vindo a ser aplicado a diferentes contextos e cenários. Áreas como Web Mining [Cavique, 2007], Marketing [Cavique e Coelho 2008], *Process Mining* [Cavique, 2015], *Financial Studies* [Marques & Cavique, 2013], [Tiple, 2014], [Tiple, Cavique & Marques, 2015], demonstram assim a sua capacidade de síntese, interpretação e eficácia na descoberta de padrões sequenciais.

3. Rede *Foursquare* e Modelo de Dados

O *Foursquare* tem sido alvo de análise por investigadores ligados ao estudo de mobilização e de redes sociais. Este capítulo focará o *Foursquare* que é a maior rede do género com aproximadamente 45 milhões de utilizadores [Foursquare, 2014].

3.1. *Foursquare*

O *Foursquare* é uma rede social baseada em localização, também conhecida como rede social móvel [Sutko & Silva, 2011], na qual o utilizador, por meio de um telemóvel, informa acerca do lugar onde está naquele instante.

A primeira versão do *Foursquare* foi desenvolvida em meados de 2008 por Dennis Crowley e Naveen Selvadurai, e teve origem num projeto anterior chamado *Dodgeball*.

Este serviço bastante semelhante ao próprio *Foursquare* foi desenvolvido na Universidade de Nova Iorque em 2000, e permitia ao utilizador o envio de mensagens escritas de telemóvel a todos os seus contactos, avisando da sua localização.

Cinco anos depois o *Dogdeball* foi adquirido pela empresa Google para a implementação do seu serviço Latitude, o que veio permitir que os utilizadores fizessem os *check-in*s nos locais e fossem classificados através de um estado de "Regular" e "VIP", dependendo do número de idas aos locais.

Em março de 2009, com o conhecimento adquirido através da criação do *Dodgeball*, Dennis Crowley e Naveen Selvadurai apresentaram a ferramenta num dos maiores eventos de inovação e tecnologia nas áreas de cinema, música e interatividade - o "South by Southwest Interactive" (SXSW), que ocorre na cidade de Austin, no estado do Texas (Estados Unidos da América).

A partir de então, com as constantes melhorias verificadas e o seu rápido crescimento, o *Foursquare* ultrapassou a categoria de rede social de georreferenciação tornando-se assim num fenómeno de expansão.

O *site* oficial define a aplicação da seguinte forma:

> "O Foursquare é uma plataforma móvel baseada em localização que torna as cidades mais fáceis e mais interessantes de explorar. Através do *check-in* os utilizadores partilham a sua localização com os seus contactos conquistando pontos e medalhas virtuais. O *Foursquare*

> é também um guia de experiências do mundo real, permitindo que os utilizadores registem locais realizando comentários e recomendações sobre os mesmos, bem como também acumular recompensas virtuais e físicas pela sua atuação." [Foursquare 2014]

O *Foursquare* é definido como a ferramenta de pesquisa mais completa e com maior número de utilizadores no mundo, aproximadamente 45 milhões em termos globais, obtendo 1 milhão de registos mundiais a cada mês.

Segundo informações do *blog* oficial, só no ano de 2010 houve um crescimento de 3400% no uso do serviço, incluindo mais de 380 milhões de *check-ins*, sendo um deles feito a partir do espaço, pela Estação Espacial Internacional - Diariamente são adicionados 35 mil novos perfis e o número de *check-ins* ultrapassa 2,5 milhões [Foursquare, 2014].

Começar a usar o *Foursquare* é um processo simples, bastando um registo no *website,* seguido da instalação do aplicativo num *smartphone* - Dispositivo Móvel de Conexão Multi-rede (DHMCM) – ou navegando através do *site*. Em ambos os casos é fundamental ter uma ligação à *internet* disponível, a qual pode ser gerada por redes Wi-Fi, EDGE, 3G, *bluetooth*, entre outras.

Após estes procedimentos, é possível adicionar utilizadores estabelecendo uma rede pessoal de contatos no serviço para, então, partilhar com os mesmos os locais onde o utilizador se encontra por meio do *check-in*. Tal como representado na Figura 3.1, o ecrã inicial - Encontrar um Lugar -, permite pesquisar os locais mais próximos da sua localização de acordo com filtros pré-sugeridos pela aplicação.

A partir da última atualização disponibilizada em Junho de 2014,o *Foursquare* passou a ter uma função apenas de pesquisa e recomendações de locais. Associada a uma nova filosofia, surge uma nova imagem e são disponibilizadas novas funcionalidades.

O objetivo passou por desagregar as principais funções, criando um aplicativo único e exclusivamente para *check-ins*, o *Swarm*, reservando-se o próprio *Foursquare* unicamente para que os utilizadores aí insiram as suas notas e opiniões sobre estabelecimentos comerciais e restaurantes.

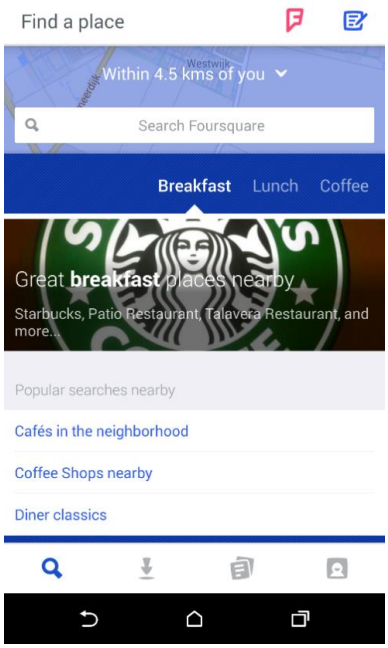

Figura 3.1 - Ecrã Inicial do *Foursquare*

Partindo da ideia de que cada pessoa explora lugares diferentemente, todos os utilizadores do *Foursquare* viram, assim, os seus *check-ins* transferidos para o *Swarm*, novo *app* que regista as visitas aos locais da aplicação.

O novo *Foursquare* parte das preferências do utilizador para as associar à geolocalização e então recomendar cafés, restaurantes e outros locais aos seus mais de 45 milhões de utilizadores.

Na página inicial passa a ser possível aceder a uma barra de navegação, na parte inferior do ecrã, onde encontramos as opções de *What's good here, Tips e User*.

De acordo com a sua localização, a opção *What's good here* (Figura 3.2) sugere uma lista de locais nos quais poderá fazer *check-in* de forma rápida e intuitiva. O índice de relevância para a indicação de cada local preferencial baseia-se no número de *check-ins* feitos e nas dicas adicionadas pelos utilizadores e confirmadas por outros participantes. Assim, é possível encontrar os locais mais populares de cada categoria

numa distância que vai de 250 metros até 10 quilómetros, com a opção de traçar a rota para chegar ao local desejado com o auxílio do *Google Maps.*

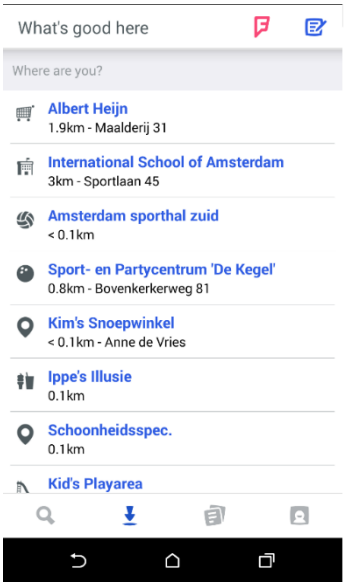

Figura 3.2 - Opção *What's good here*

O separador *Tips* (Figura 3.3) contém dicas de locais que estão classificados de acordo com os gostos de cada perfil.

O *Foursquare* faculta a possibilidade de visualizar todas as dicas e permite aos utilizadores organizá-las de forma a que se possam pesquisar locais recomendados de acordo com o raio de distância onde o utilizador se encontra; permite ainda que as categorias sejam organizadas segundo as que têm mais preferência e recompensas físicas.

Para cada dica é possível Gostar ou Guardar, o que equivale a dizer que se concorda com esse comentário ou se pretende visitar o local para confirmar essa mesma dica. A opção de Guardar permite criar uma lista pessoal de tarefas, com uma relação de locais para visitar, dicas para publicar, o que usar em certos espaços, que compras fazer e outras opções.

Figura 3.3 - Opção *Tips*

No ecrã do utilizador (Figura 3.4) é possível visualizar um resumo do perfil com o número de seguidores e o número de dicas que registadas na aplicação. Cada utilizador poderá ter acesso ao número de vezes que cada dica foi visualizada por outros utilizadores e configurar os seus gostos, os quais irão ter um impacto significativo no comportamento da aplicação.

O *Foursquare* tem vindo a renovar-se e trouxe o estímulo da competição ao introduzir um sistema de pontuação para os *check-in*s.

Um dos principais fatores de sucesso do *Foursquare* está relacionado com o facto de disponibilizar serviços de *gamificação*, em que o uso de conceitos normalmente aplicados a jogos é também utilizado em aplicações [Zichermann & Linder, 2010].

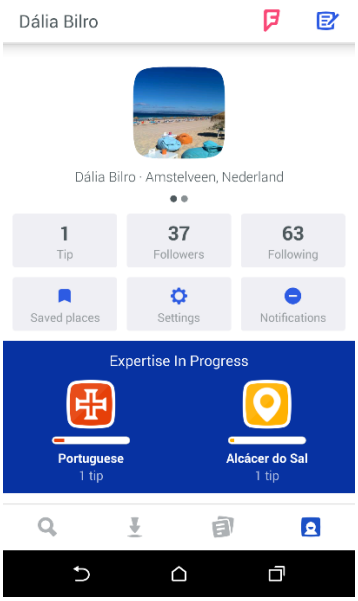

Figura 3.4 - Histórico e configuração do utilizador

Segundo a pesquisa de Zago e Recuero [2011], um outro indicativo forte de motivação reside no facto de o *Foursquare* possuir uma dinâmica lúdica que incentiva a criação e manutenção de esferas íntimas, onde as interações entre utilizadores acabam por ser partilhadas entre circuitos fechados. Os estudos de [Sutko & Silva, 2011] mostram que, devido ao seu sistema de autenticação com perfis personalizáveis, as redes sociais móveis induzem um envolvimento apenas para com semelhantes – uma espécie de autenticação social.

Para além dos benefícios oferecidos pelo jogo e do carácter lúdico da aplicação, o *Foursquare* também disponibiliza um sistema exclusivo mais direcionado para as empresas; mediante um registo, estas poderão usar o serviço como ferramenta de *marketing,* possibilitando gerir e criar as suas próprias estratégias. É neste âmbito que o *Foursquare* demonstra ser bem mais que uma plataforma social, uma vez que é capaz de se transformar numa plataforma de negócios com capacidade de gerar uma experiência diferenciada e inovadora aos consumidores que fazem usam desta tecnologia. Através de um baixo investimento e em virtude das experiências vividas através da aplicação, é possível implementar estratégias para recompensar, fidelizar e prospetar os clientes a uma determinada empresa ou marca.

Finalmente, poderá apontar-se como fator de motivação o facto de o serviço estar disponível para várias plataformas móveis e poder ser acedido pelo navegador de telemóveis e computadores, aspetos que favorecem o seu crescimento, principalmente no âmbito das redes sociais.

3.2. Princípio Básico: *Check-in*

O *check-in* é o grande ponto de partida para o uso do *Foursquare*. É através desta função que os utilizadores partilham com os restantes contactos a sua localização. Realizar um *check-in* equivale a marcar uma presença física num determinado lugar.

Ao realizar o *check-in*, o utilizador ganha pontos, os quais serão utilizados para a criação de *rankings* automáticos entre a sua rede de contactos.

Na última atualização disponibilizada pelo *Foursquare*, como já referido, o *check-in* passa a ser gerido por uma aplicação externa, o *Swarm*. Quando, através do *Foursquare*, o utilizador inicia o processo de *check-in* num determinado local (Figura 3.5), a aplicação do *Swarm* surge automaticamente.

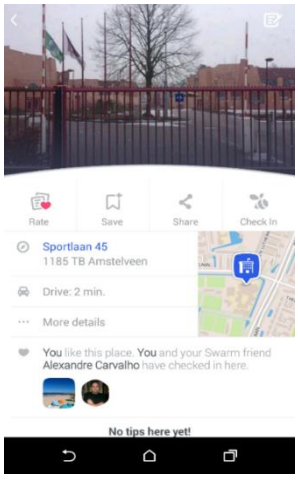

Figura 3.5 - Opções disponíveis para um local

No momento da realização do *check-in* (Figura 3.6), para além de uma breve descrição feita pelo utilizador, é também possível adicionar uma fotografia e partilhar essa informação nas restantes redes sociais.

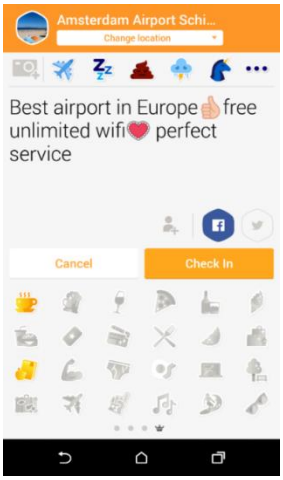

Figura 3.6 - Menu de *Check-in*

Caso o local em que se pretende fazer o *check-in* não esteja listado, o utilizador pode ainda assim fazê-lo, pois tem a hipótese de o criar, dando até a conhecer alguma informação complementar (morada, telefone, *site*, *email*, etc),

Cada local no *Foursquare* é classificado segundo uma das nove categorias pré-definidas: Arte & Entretenimento, Escolas e Universidades, Comida, Espaços Exteriores, Lugares Noturnos, Viagens, Residência, Lojas/Serviços e Outros Serviços.

A conta do *Foursquare* pode ser configurada a partir da conta do *Twitter* e/ou do *Facebook*, o que permite que a informação da partilha do *check-in* seja maximizada para os contactos dessas redes sociais. Além disso, é possível receber notificações sobre a sua rede de contactos que realizaram *check-ins* nas suas proximidades.

No entanto, não é só para partilhar as suas localizações que os utilizadores fazem *check-ins*. Existe a opção *off–the-grid*, na qual a localização não é divulgada, mas a contabilização de pontos é igualmente efetuada.

O sistema também oferece medalhas virtuais (*Golden Stickers*), *mayorship* (*status* ganho por quem mais frequenta determinado local) e recompensas (caso o estabelecimento ofereça alguma vantagem para que se realize o *check-in* no local).

O utilizador, ao encontrar um determinado local no *Foursquare*, tem a possibilidade de saber quem são os utilizadores que o frequentaram e quais são as dicas e informações relativas a esse mesmo local. Existe ainda a possibilidade de marcar numa lista (*to-do list*) o que o utilizador pretende fazer/comprar/usar em determinado local, com base nas dicas deixadas pelos restantes utilizadores.

3.3. Medalhas Virtuais (*Golden Stickers*)

A atribuição de medalhas virtuais é uma maneira de recompensar os utilizadores pela utilização do *Foursquare*.

Um dos principais incentivos para que se utilize o *Foursquare* está relacionado com o conceito de jogo que subjaz na aplicação; assim, ganha mais medalhas quem mais participa na aplicação.

A primeira medalha virtual é atribuída automaticamente, no momento em que é feito o primeiro *check-in*; uma segunda medalha poderá ser obtida a partir do instante em que se realizem três *check-in*s no mesmo lugar, na mesma semana.

Para além destas, existe ainda a hipótese de obter outras medalhas para os utilizadores que realizem *check-in*s em dez lugares diferentes. Muitas outras são possíveis de obter, de acordo com a participação do utilizador e com a própria popularidade que o *Foursquare* vai ganhando nas redes sociais. A tendência é para que quanto mais pessoas usarem o *Foursquare*, mais medalhas virtuais vão sendo criadas para cada público-alvo. Por exemplo, Nova Iorque e São Francisco possuem medalhas específicas para empresas como a *Starbucks* e a MTV.

Para se tornar *Mayor* (Presidente) de um local, é preciso ter feito *check-in* nesse lugar em mais dias do que qualquer outro utilizador nos últimos dois meses (60 dias). Para garantir a *mayorship* de um local é preciso realizar *check-in*s constantes, sob risco de que outros utilizadores possam mesmo acabar por ficar com o título.

Desse modo, ser *Mayor* de um local pode indicar também uma relação de fidelidade, principalmente no caso de estabelecimentos comerciais.

Muitas empresas já estão mais atentas para este facto e estão a recompensar fisicamente os seus *Mayors* através de diferentes formas - os *specials*.

O serviço incluiu também a função Radar, que avisa quando há algo interessante nas proximidades do local onde o utilizador se encontra.

3.4. Caracterização do *Dataset Foursquare*

O *Foursquare* é uma aplicação que permite aos utilizadores partilhar com os seus contactos a sua localização, através de *check-ins*. Os milhões de *check-ins* efetuados diariamente pelos seus mais de 45 milhões de utilizadores representam 60 milhões de locais registados na aplicação.

Na figura 3.7 podemos observar o modelo relacional que organiza os dados em forma de tabelas e permite também definir relações entre elas:

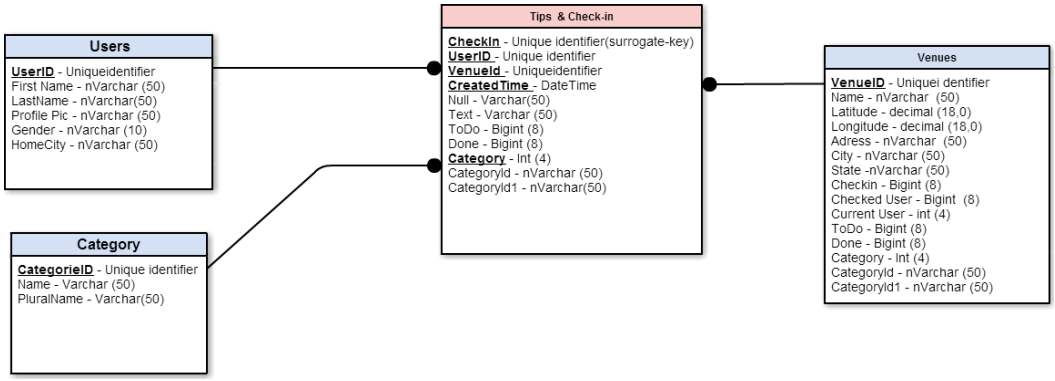

Figura 3.7 - Modelo Relacional do *dataset Foursquare*

A tabela *Users* permite armazenar toda a informação que diz respeito aos utilizadores da aplicação. A tabela *Venues* identifica os locais e todas as suas características, com forte desataque para os campos Latitude e Longitude que terão um

papel fundamental na fase de redução dos dados. As nove Categorias existentes referentes aos vários locais do *dataset* estão guardadas na tabela *Category*.

Por fim, a tabela *Tips & Check-in* que assume um papel bastante relevante neste estudo pois reúne todos os comentários de todos os utilizadores referente aos locais existentes na aplicação. É através desta tabela que serão extraídos os *check-ins* da aplicação, identificando o ID do local, o ID do utilizador, a data e o comentário em relação a esse local. Para tal foi necessário criar uma chave substituta (*surrogate key*) com o nome de *Check-in* que serve como identificador único para cada registro na tabela e que é resultado da concatenação dos atributos UserID, VenueID e CreatedTime.

Aqui estão reunidos os *check-ins* compreendidos entre 20 outubro de 2008 e 4 de outubro de 2011.

A tabela 3.1 mostra o número utilizadores, lugares, *check-in* e categorias do *dataset Foursquare*:

Tabela 3.1 – *Dataset Foursquare*

Nº. utilizadores	**Nº. lugares**	**Nº. *check-ins***	**Nº. Categorias**
49.001	157.655	315.852	9

Uma análise detalhada das Estatística Descritiva do *dataset* e a Transformação de dados é realizado no capítulo seguinte.

4. Análise e Transformação dos Dados

Este capítulo, para além de abordar a metodologia utilizada neste estudo, apresenta uma análise estatística do *dataset Foursquare,* de forma obter-se um entendimento básico dos seus dados e das relações existentes entre as variáveis analisadas. O maior enfoque será dado à fase de redução de dados do *dataset,* a qual se tornou um fator bastante importante nos resultados posteriormente obtidos.

4.1. Metodologia

Este projeto foi desenvolvido utilizando as ferramentas *SAS Enterprise Guide* V.6.1 e *Enterprise Miner* V. 13.1. Este *software* é utilizado em todos os setores de atividade e tem a vantagem de poder ser instalado em diversos ambientes operacionais disponíveis no mercado, sendo ainda capaz de tratar quaisquer dados, qualquer que seja a sua origem.

O *Enterprise Miner* é uma solução integrada que, através de uma interface totalmente gráfico, possibilita ao utilizador o acesso a diferentes etapas da metodologia *SAS Institute* para o *Data Mining*: a metodologia SEMMA. Esta metodologia decompõe-se em 5 etapas: *Sample; Explore; Modify; Model* e *Assess*.

De uma forma sucinta é possível apresentar esta metodologia como um processo de 5 fases, que se inicia com uma amostra (*Sample*) representativa dos dados, à qual se aplicam técnicas estatísticas de exploração e de visualização dos dados (*Explore*). Posteriormente são selecionadas e transformadas as variáveis (*Modify*) consideradas mais significativas (as variáveis que sobressaíram na fase anterior), as que são mais relevantes em termos de projeto, e sobre as quais se constroem os modelos (*Model*); por fim avalia-se o modelo (*Assess*). Cada uma das etapas é distinta e corresponde a um ciclo, e as suas tarefas internas podem ser executadas repetidamente sempre que necessário, i.e, pode-se atualizar e ajustar sempre que surgir nova informação [SAS, 2014].

A interface gráfica permite construir diagramas de projetos de *Data Mining*. Estes diagramas são formados por nós que executam operações especializadas sobre os

dados. Os nodos são interligados por setas que o utilizador utiliza para definir qual a sequência de operações que pretende executar no processo de *Data Mining*.

O *SAS Enterprise Guide* é "uma ferramenta OLAP para Windows, orientado por projetos, e que possibilita o acesso rápido a uma grande parte da potencialidade analítica do SAS para estatísticos, analistas de negócios e programadores SAS" [SAS, 2014].

Nesta ferramenta efetuaram-se os primeiros trabalhos de pré-processamento e análise descritiva dos dados, como explicado no subcapítulo seguinte.

A Descoberta de Conhecimento em Bases de Dados (DCBD) utilizada em vários tipos de aplicações tem o apoio de técnicas e ferramentas de grande utilidade para a sua realização. Porém, a adoção de uma metodologia também é de fundamental importância para se tentar estabelecer um planeamento e organização da execução do processo de DCBD.

Das várias metodologias de DCBD disponíveis, a aplicada nesta fase do trabalho foi a metodologia defendida por [Fayyad et al., 1996], com a conjugação da metodologia embebida no *software* utilizado – SEMMA.

Atendendo ao facto de as metodologias partilharem a mesma essência, o caso prático apresentado neste trabalho foi desenvolvido segundo uma metodologia própria, já que a metodologia descrita por Fayad, embora mais completa, tem uma visão mais ampla e a SEMMA coloca uma ênfase maior na recolha da amostra.

Segundo Fayyad et al. [1996], a definição da Descoberta de Conhecimento em Bases de Dados (DCBD) é dada como o processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis implícitos nos dados.

A Figura 4.1 mostra uma visão geral das Fases Inclusas no Processo de DCBD a partir do momento em que os objetivos estão definidos. O processo normalmente não é linear, e envolve uma forte interação com o utilizador e várias iterações entre as suas fases constituintes. Essas iterações estão representadas, na figura, pelas setas a tracejado, e podem inclusivamente fazer o processo voltar à fase de especificação dos objetivos.

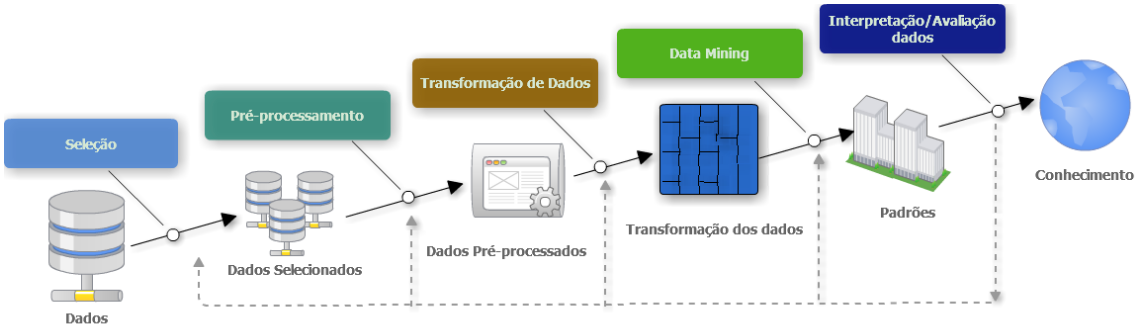

Figura 4.1 - Processo de DCBD [Fayyad et al., 1996]

4.2. Variáveis do problema

De uma forma geral, a fase de Pré-Processamento engloba uma análise inicial dos dados para se obterem sólidas definições dos mesmos, e toda e qualquer operação necessária para a escolha dos dados mais relevantes face aos objetivos do utilizador.

A etapa de pré-processamento dos dados é a fase que envolve um elevado nível de conhecimento e é reconhecida como sendo a fase mais demorada, chegando a consumir cerca de 80% do tempo total.

Ressalta-se que, assim como no processo geral de DCBD, não existe uma sequência obrigatória quanto à ocorrência das subfases de pré-processamento, pois dependendo da situação pode-se, por exemplo, preferir realizar a transformação dos dados antes mesmo de uma determinada limpeza de dados. Os trabalhos de pré-processamento seguintes foram desenvolvidos em *SAS Guide*, e *SAS Miner*, conforme as especificidades do problema.

4.2.1. Preparação da base de dados do trabalho

Uma base de dados concebida corretamente fornece acesso a informações atualizadas e precisas, sendo essencial investir algum tempo para cumprir os princípios de conceção de uma estrutura adequada. No *dataset* em questão, existiu a necessidade de preparar os dados de forma a garantir essa mesma estrutura.

No geral, a informação disponível nos quatros ficheiros era bastante percetível. O ficheiro que sofreu um maior nível de tratamento foi o das *Tips & Check-ins*, no qual teve que se definir um método para ler os dados e modificar toda a sua estrutura. A questão que existia era a de que o ficheiro estava organizado por *UserId* e, conforme o número de dicas que esse utilizador partilhava, o registo era sempre acrescentado no final dessa mesma linha, o que resultava num ficheiro com 480 colunas, onde os registos ficavam sempre separados por *VenueId*, seguidos de um campo NULL.

Através desse critério e com recurso à linguagem SAS foi criado um procedimento de autoria própria que visa a importação dos dados, de forma a integrar os dados num novo ficheiro.

O algoritmo, após efetuar a leitura dos dados, executa uma macro onde as variáveis de argumento são a coluna inicial e a coluna final e percorre, através de um ciclo do *While,* todas as 480 colunas. Ao mesmo tempo é criada uma *View* temporária para guardar a cadeia de registos, onde a variável em estudo é igual a *Null*. Essa informação é guardada numa tabela final e a *View* temporária é apagada para guardar o próximo registo.

A figura 4.2, que representa a tabela final, permite-nos visualizar o ficheiro após a aplicação do procedimento de transformação.

user_id	local_id	nulo	comentario	data_unix	tod	do	ncate	category_1
33965	4c613eb412e5c9287809214c	null	BaconpackedBLT,can'tbeat	1297297681	0	1	1	4bf58dd8d48988d16f941735
4104366	4bd70ea04e32d13a4d96c380	null	Notcominghereagaintheych	1300238358	0	1	1	4bf58dd8d48988d1d2941735
2429331	4bb226ddf964a520e1bd3ce3	null	breakfastsandwichesforali	1280761502	1	2	1	4bf58dd8d48988d146941735
3041547	4ca4311 6f88a9521fe0ab049	null	KennyPowersautographsignin	1285829012	0	1	0	
2636583	4defc0ce45dd29e268a6d29f	null	DanBisgoingDOWWWWWNNNNNNN	1308446358	0	1	0	

Figura 4.2 - Exemplo do registo após transformação

O principal objetivo desta tarefa foi o de transformar os dados, o que envolveu a aplicação de técnicas sobre as variáveis, de forma a maximizar a informação contida nas mesmas, tornando-a assim mais apropriada para a modelação.

4.2.2. Seleção e limpeza de dados

A seleção de dados envolve a escolha das tabelas, dos seus atributos e das suas instâncias em relação aos objetivos do utilizador, considerando-se ainda que, em caso de necessidade de manipulação de informações de várias tabelas, se proceda à sua integração de modo a que se obtenha um conjunto único de instâncias sobre o qual será dada a continuidade do pré-processamento e/ou do processo DCBD.

A Figura 4.3 descreve o processo de integração das várias tabelas utilizando o *Enterprise Guide*.

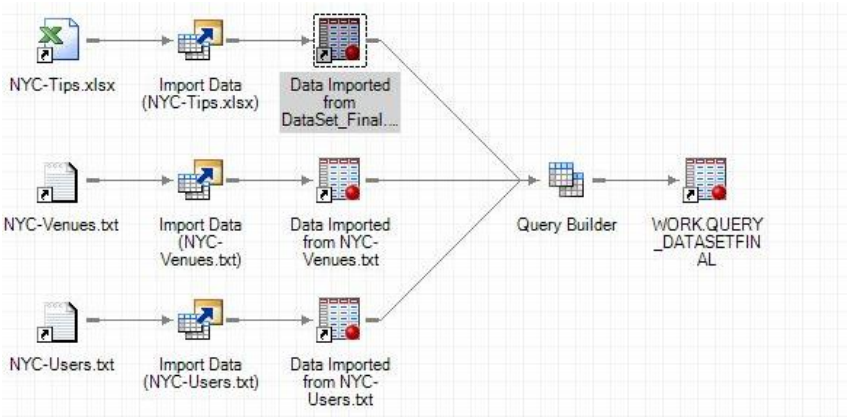

Figura 4.3 - Processo de integração das tabelas

Com recurso a uma *query* que reúne o conteúdo das três tabelas, foi possível escolher os atributos a considerar. Nesta fase da seleção, os dados escolhidos estão diretamente relacionados com os utilizadores e/ou locais visitados e podem ser alvo de análise tal como representado e explicado na Figura 4.4.

A etapa da limpeza tem como objetivo aumentar a qualidade dos dados selecionados. De início, utilizou-se o *dataset* de visitas de Nova Iorque, registadas entre o período de 20 de outubro de 2008 e 4 de outubro de 2011. Entretanto, observou-se que o ficheiro apresentava muitos registos que não respeitavam esse critério, pois continham essa informação para localizações diferentes.

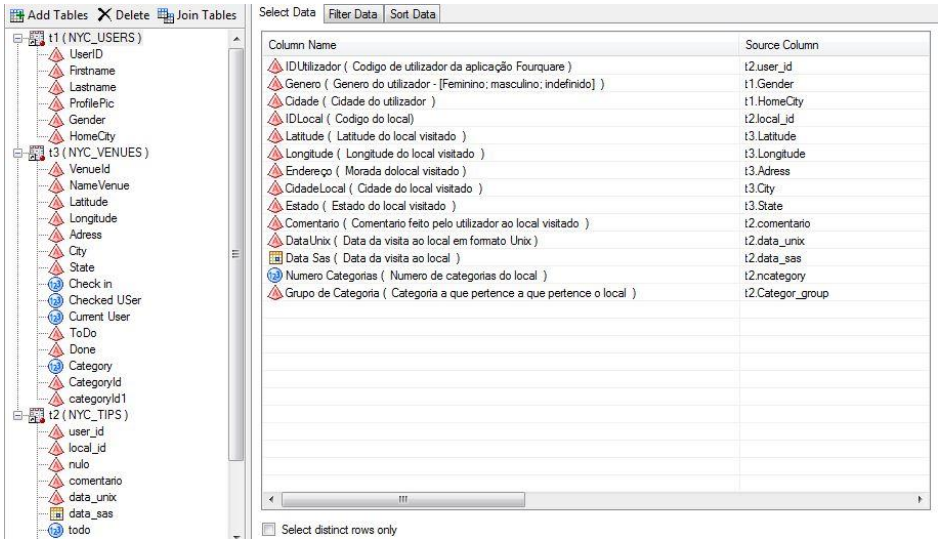

Figura 4.4 - Escolha e identificação de atributos

Por se tratar de zonas completamente distantes, tal impossibilitava a análise gráfica e o estudo de padrões das sequências efetuadas por cada utilizador. A solução adotada passou pela exclusão dos registos com estas irregularidades; seguidamente, os registos passaram a ser filtrados através de coordenadas de latitude/longitude (Figura 4.5), correspondentes ao estado de Nova Iorque.

```
PROC SQL NOEXEC;
    SELECT /* USER_ID */
             (INPUT(t1.user_id,comma12.)) AS USER_ID,
           t1.Gender,
           t1.HomeCity,
           t1.local_id,
           /* LATITUDE */
             (INPUT(t1.Latitude,comma12.)) AS LATITUDE,
           /* LONGITUDE */
             (INPUT(t1.Longitude,comma12.)) AS LONGITUDE,
           t1.Adress,
           t1.City,
           t1.State,
           t1.'Check in'n,
           t1.'Checked USer'n,
           t1.'Current User'n,
           t1.comentario,
           t1.nulo,
           t1.data_unix,
           t1.data_sas,
           t1.todo,
           t1.done,
           t1.ncategory,
           t1.category_1,
           t1.category_2,
           t1.Categor_group
      FROM SASUSER.QUERY_DATASETFINAL t1
      WHERE (CALCULATED LATITUDE) BETWEEN 40.47739900000000 AND 40.91757700000000 AND (CALCULATED LONGITUDE) BETWEEN
           -74.25909000000000 AND -73.70027200000000;
QUIT;
```

Figura 4.5 - Filtro de registos através das coordenadas Latitude e Longitude

Na Tabela 4.1 podemos verificar o número de registos existentes antes e depois da limpeza de dados. Observou-se uma redução bastante significativa no volume de dados do *dataset*, sendo que os lugares diminuíram 63%, e cerca de 42% no total do número de visitas.

Tabela 4.1 - Análise comparativa do número de registos antes e depois da limpeza dos dados

	Número de registos	
	Antes	**Depois**
Users	49.001	48.308
Venues	157.655	56.854
Check-in	315.852	183.211
Category	9	9

4.2.3. Análise de *outliers* e *missing values*

Um exemplo que é bastante comum na limpeza dos dados (e que ocorre no pré-processamento) é a procura por valores absurdos (deteção de *outliers*), valores esses que à partida não deveriam existir na base de dados por serem impossíveis.

Muitas vezes, nesta etapa, é necessário recorrer a uma análise para além dos dados, pois valores que, numa primeira análise, parecem valores irreais, podem de facto ser valores importantes no processo e capazes de prover a análise com *inputs* importantes para a compreensão de certo tipo de comportamentos.

Uma das técnicas usadas para identificação de *outliers* é feita através da interpretação do *Box Plot* das variáveis, calculando-se a mediana, o quartil inferior (Q1) e o quartil superior (Q3). A subtração entre o quartil inferior (Q1) e o quartil superior (Q3) é igual a (L).

Os valores que forem maiores que Q3+3L e menores que Q1-3L devem ser considerados suspeitos de pertencer à população, devendo ser investigada a origem da dispersão. Tal como iremos verificar no subcapítulo seguinte, existem variáveis nesta

situação; no entanto, optou-se por manter esses registos devido ao facto do conjunto de valores se mostrar relevante para a construção de sequências.

Não foram identificados *missing values*, não só nas variáveis Intervalares como também nas variáveis de Classe.

4.2.4. Redução de Dados

A primeira fase da metodologia SEMMA consiste na obtenção de uma amostra significativa a partir da extração de uma quantidade de dados do universo existente. A amostra deve corresponder a um subconjunto de dados que pertencem ao universo onde cada elemento tem as mesmas hipóteses de ser incluído, mas também deve ser pequena de modo a tornar-se rápida e de fácil manipulação. Neste caso, o volume de dados era suficientemente grande para tornar o processo de análise dos dados impraticáveis.

Na tabela 4.2 podemos observar os rácios das variáveis *UserId* e *VenueId* do *dataset*. Deve-se ressaltar que o número médio de *check-ins*/*utilizador* é bastante inferior ao apresentado em outros estudos [Cheng et al., 2011].

O rácio de *check-in* por local é de 3,22 e o rácio de *check-in* por utilizador é de 3,79. Os valores apresentados representam rácios baixos o que poderá indicar que os utilizadores não são utilizadores frequentes da aplicação e que o número de locais visitados por esses utilizadores também não é elevado. Na prática poderá também indicar um baixo número de sequências encontradas, o que levará de certa forma a apresentar uma alternativa para a redução de dados em subconjuntos com vista a estudar diferentes abordagens.

Tabela 4.2 - Rácios das variáveis *UserId* e *VenueId*

Rácios	**Utilizadores** **48.308**	**Locais** **56.854**	***Nº. check-ins*** **183.211**
Utilizadores	---		
Locais	1,17	---	
Nº. check-ins	3,79	3,22	---

4.2.5. *Check-ins*, utilizadores e locais

A fase de compreensão dos dados inclui a identificação da informação relevante para o estudo e uma primeira familiarização com o conteúdo, descrição, qualidade e utilidade dos dados.

O primeiro passo na etapa de reconhecimento de uma base de dados (que será efetuado em *SAS Guide*) é analisar as medidas descritivas da mesma, tanto da base como do seu todo.

Esta etapa é fundamental, pois uma análise descritiva detalhada permite a familiarização com os dados, organizá-los e sintetizá-los, de forma a obter as informações necessárias do seu conjunto e a responder às questões que estão a ser estudadas.

Nesta fase, as análises gráficas desempenham um papel bastante importante, uma vez que são as mais indicadas em situações cujo objetivo é dar uma visão mais rápida e fácil das variáveis às quais se referem os dados.

Em seguida, através da figura 4.6, apresenta-se um quadro com o resumo geral das variáveis; as métricas utilizadas nesta análise foram a média, o desvio padrão, o mínimo, o máximo, a quantidade de registos e os registos nulos. De destacar que não existem *missing values* em nenhuma das variáveis aqui representadas.

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	N Miss
LOCALID	LocalId	17882.75	16633.25	1.0000000	56854.00	183211	0
LATITUDE	Latitude	40.7391616	0.0541438	40.4835150	40.9175280	183211	0
LONGITUDE	Longitude	-73.9809047	0.0746953	-74.2590380	-73.7002798	183211	0
USER_ID	UserId	1974842.80	2346232.63	32.0000000	10880575.00	183211	0
Check in	Checkin	3233.56	11912.29	0	162134.00	183211	0
ncategory		1.5289966	1.0787473	0	14.0000000	183211	0

Generated by the SAS System ('Local', X64_7PRO) on July 14, 2015 at 5:55:54 PM

Figura 4.6 - Sumário Estatístico

O histograma da Figura 4.7 apresenta-nos o comportamento da variável numérica *UserId*, que representa o código unívoco de cada utilizador da aplicação. No eixo x estão representados esses mesmos códigos e no eixo de coordenadas y, a sua frequência relativa. É possível observar uma concentração de percentagem de observações à esquerda do eixo x, o que poderá apontar valores máximos bastante discrepantes, podendo significar a presença de *outliers*.

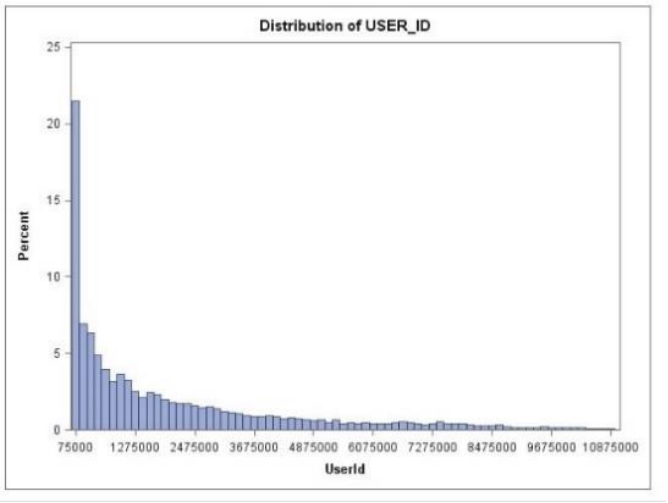

Figura 4.7 - Histograma da variável *UserId*

O mesmo comportamento se pode verificar para a variável LocalId, que identifica cada local a visitar (Figura 4.8). A possível presença de *outliers* nestas duas variáveis é um fator que terá que se ter em conta na etapa de pré-processamento

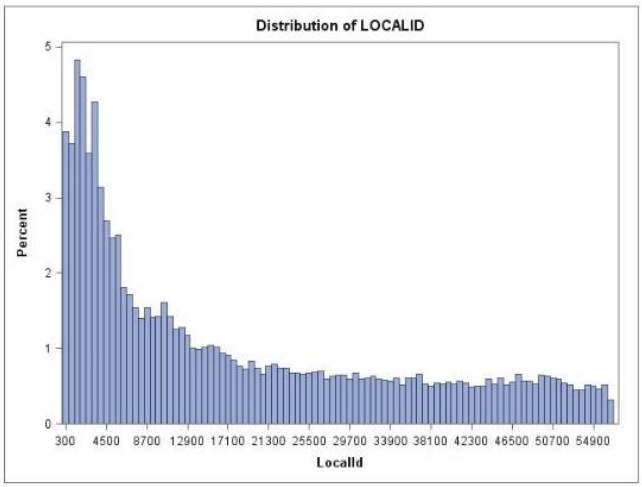

Figura 4.8 - Distribuição da variável LocalId

Baseado no pressuposto de que o estudo a realizar irá essencialmente incidir sobre padrões de locais, o mesmo só irá ser possível com recurso a sequências completas de utilizadores referente aos locais visitados.

Avaliando a distribuição das percentagens das variáveis *UserId* e *LocalId*, através da Figura 4.9 e 4.10, pode brevemente concluir-se que:

- Mais de metade dos locais apenas contém um *check-in* registado no *dataset*;
- Apenas15% dos locais foram visitados mais de 4 vezes pelos utilizadores;
- 38 utilizadores em cada 100 só efetuaram um *check-in* na aplicação;
- Mais de metade dos utilizadores efetuaram pelo menos dois *check-ins* na aplicação.

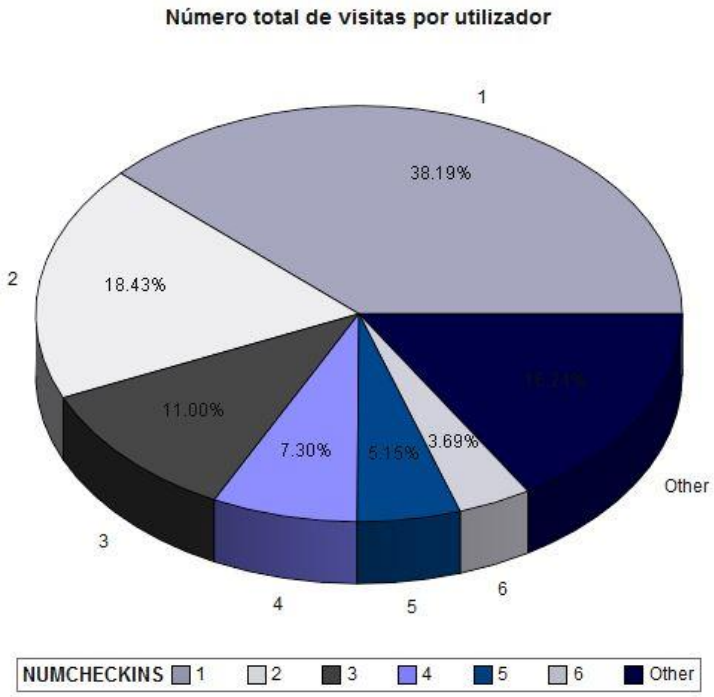

Figura 4.9 - Percentagem de número de visitas por utilizador

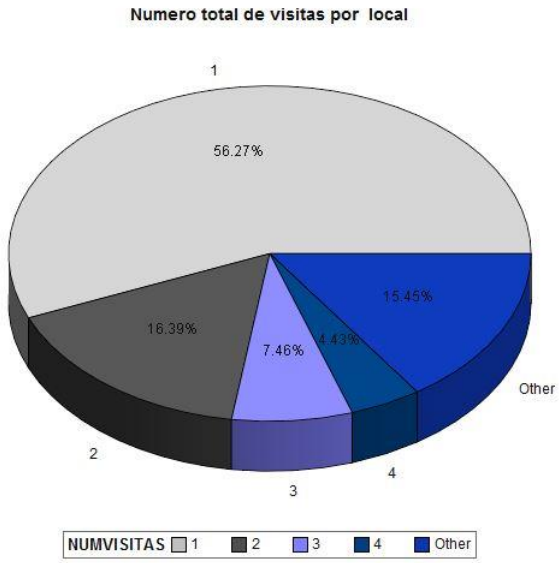

Figura 4.10 - Percentagem de número de visitas por local

4.2.6. Análise Complementar

Quanto às variáveis Latitude e Longitude, ambas apresentam uma distribuição normal - a frequência mais alta está representada no centro e decresce gradualmente para as caudas, de maneira simétrica e em forma de sino (Figura 4.11). A média e a mediana são aproximadamente iguais e localizam-se no ponto de pico, ou seja, no centro do histograma.

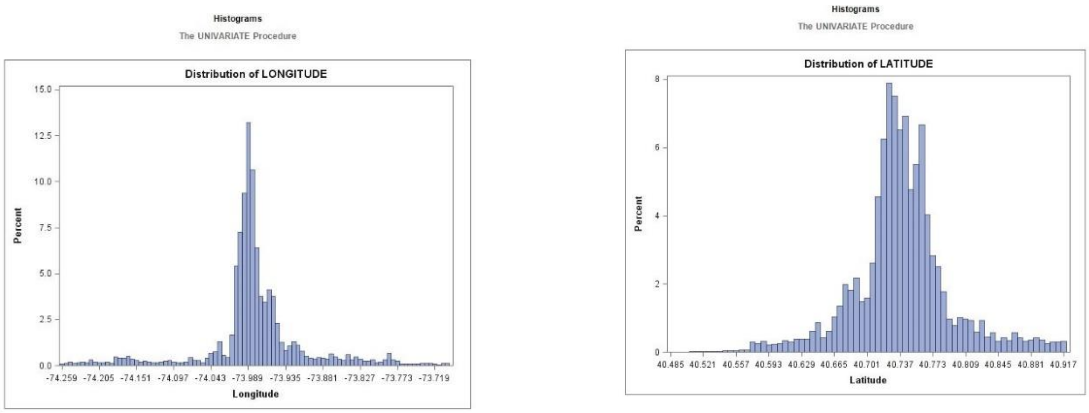

Figura 4.11 - Histograma das variáveis Latitude e Longitude

Este tipo de distribuição poderá ser explicado por existir uma grande concentração de *check*-ins nalgumas zonas muito específicas. Tal como mostra a Figura 4.12, tal facto acabou por se verificar por grande parte dos *check-ins* terem sido efetuados essencialmente em três grandes zonas (*county*) de New York, as quais estão representadas a azul escuro - Bronx, New York e Chenango.

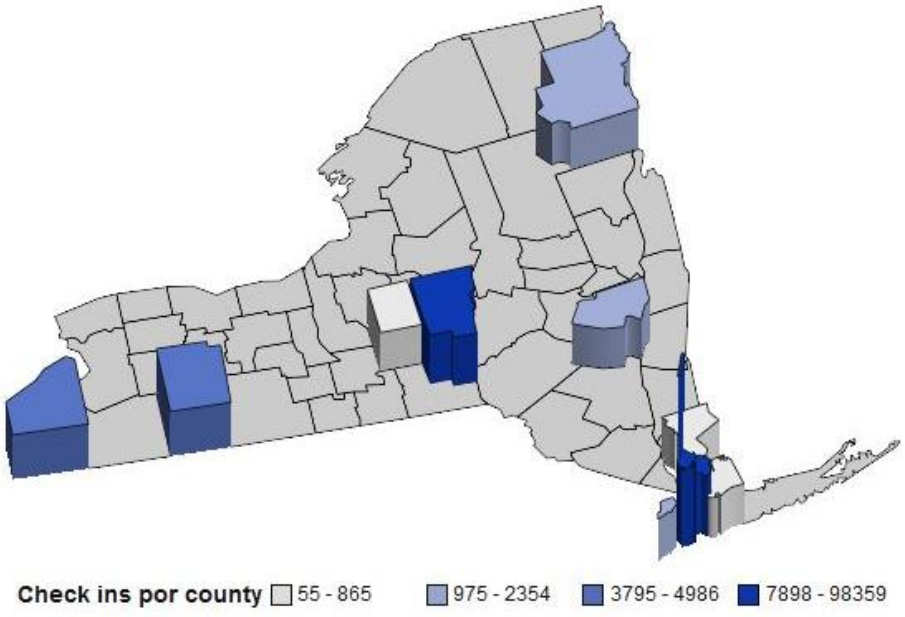

Figura 4.12 - Distribuição de *check-ins* por *county*

A variável *Data* representa a data e hora de cada visita registada na aplicação. Pela observação do seu histograma e *boxplot* (Figura 4.13), podemos verificar que grande parte dos *check-ins* foram efetuados entre o período de julho de 2010 e julho de 2011, o que se pode considerar um comportamento normal devido ao facto de, no ano de 2011, a aplicação ter tido um crescimento de 300% face ao ano anterior [Foursquare, 2014].

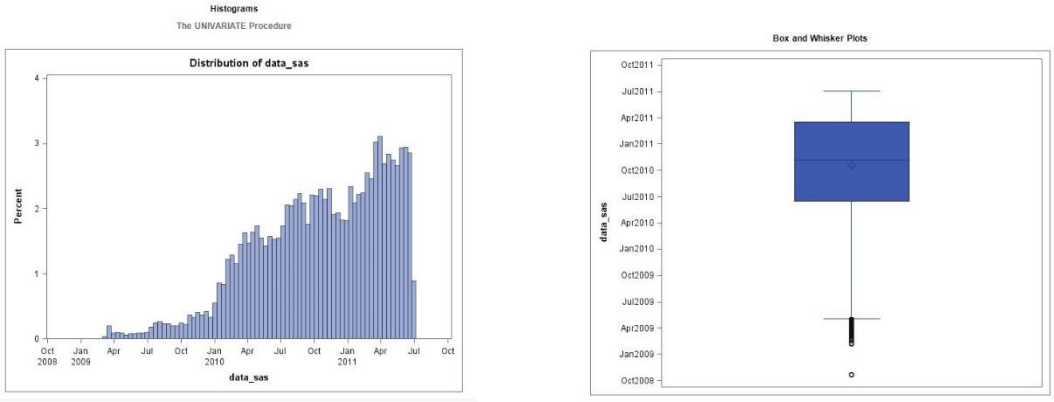

Figura 4.13 - Histograma e *Bloxpot* da variável *Data*

Na Figura 4.14 observa-se que há um padrão de comportamento no que diz respeito aos *check-ins* feitos durante a semana, pois estão distribuídos de forma equitativa e existe apenas uma pequena diferença em relação ao fim de semana.

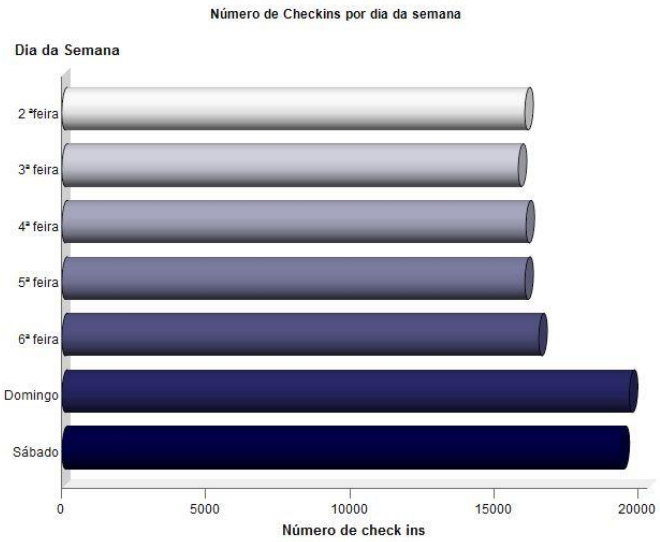

Figura 4.14 - Distribuição de *check-ins* por dia da semana

Analisando com mais detalhe esses dados, a Figura 4.15 mostra-nos as horas a que foram realizados os *check-ins* em cada dia da semana e podemos constatar, assim, que existe um acréscimo no número de *check-ins* nos dias da semana, conforme a

semana vai progredindo. O mesmo comportamento é visível em relação às horas do dia, atingindo o seu maior registo perto da hora de jantar. Este facto pode ser explicado recorrendo à análise do gráfico das categorias, uma vez que cerca de 40% dos locais pertencem à categoria *Food*. Como o jantar é a refeição mais importante do dia para os nova-iorquinos, os restaurantes abrem cedo e a maior parte, inclusivamente, abre apenas à hora de jantar.

As transições entre os dias podem ser observadas pela redução de atividade a partir do período das 2h da manhã, sendo esta é retomada no intervalo entre as 8h e 10h da manhã.

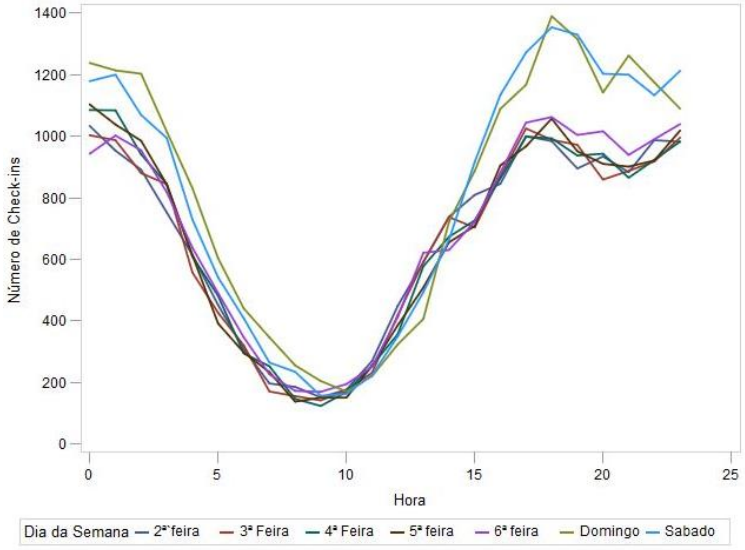

Figura 4.15 - *Check*-ins por hora

Através desta primeira análise descritiva feita à variável Categoria (Figura 4.16), na qual se analisa o número total de *check-ins* para cada categoria sobre o conjunto total de dados, observa-se, de novo, que a categoria com maior número de *check-ins* é a categoria *Food*.

A categoria *Home and Work* apresenta uma percentagem significativa de *check-ins*, o que poderá indicar que as pessoas usam o serviço não apenas quando saem à

noite (para dar a conhecer, pela sua rede social, onde se encontram nesse preciso momento), mas ainda quando saem e/ou entram em casa.

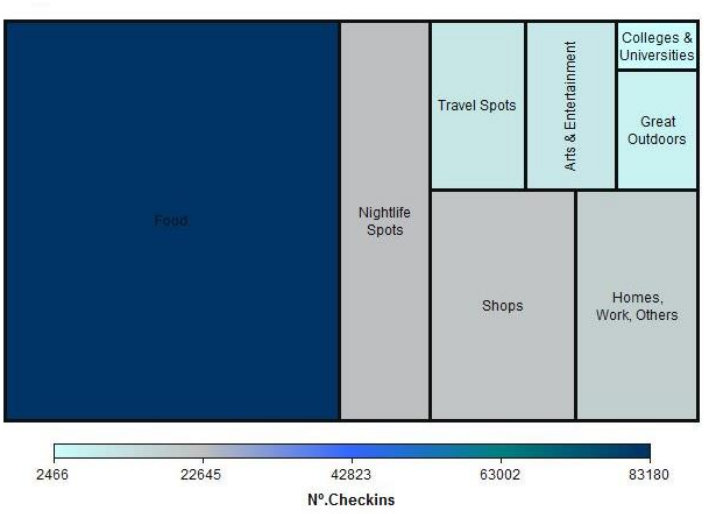

Figura 4.16 - Total de *check-ins* por categoria

Efetuou-se uma análise mais detalhada da distribuição dos *check-ins* efetuados em *Home and Work* e *NigthLife Spots* (Figura 4.17 e Figura 4.18), cujos resultados corresponderam ao esperado. Na categoria *Home and Work* observou-se que a linha começa a crescer a partir das 9 horas, coincidindo com o início do dia de trabalho. Ao sábado e domingo verifica-se um registo inferior, pois é fim-de-semana.

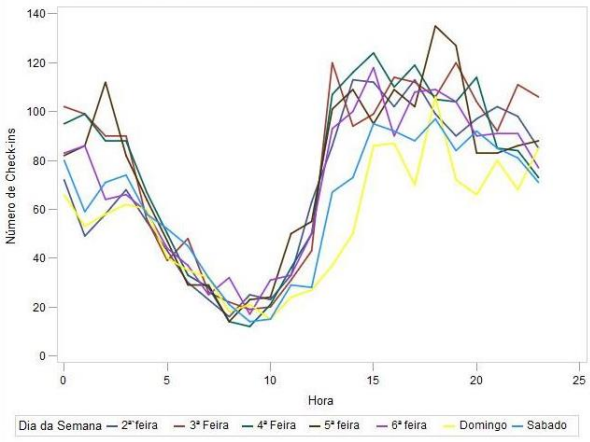

Figura 4.17 - Distribuição de *check-ins* da categoria *Home and Work* por dia da semana

Em relação aos *check-ins* efetuados na categoria *NigthLife Spots*, o pico dos registos encontra-se nas horas mais tardias, uma vez que os utilizadores fazem *check-in* em locais de diversão noturna. Denota-se, depois, um decréscimo durante o período da madrugada. As linhas a amarelo e azul representam os dias do fim-de-semana, ligeiramente superiores, porquanto são os dias propícios para a frequência de locais de diversão noturna.

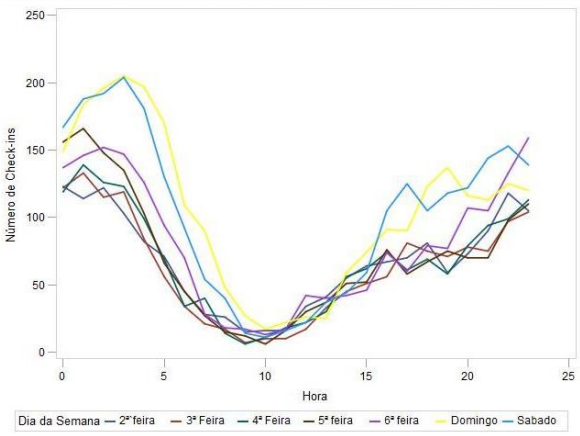

Figura 4.18 - Distribuição *Check-ins* da categoria *NigthLife Spots* por dia da semana

Quanto ao perfil dos utilizadores deste *dataset Foursquare*, podemos concluir que, maioritariamente, não são utilizadores frequentes da aplicação, pois existe uma baixa percentagem com mais de 6 *check-ins* efetuados. Os locais mais visitados encontram-se distribuídos essencialmente por 3 zonas e a grande maioria foi visitada, pelo menos, mais de 2 vezes.

Um facto com algum interesse nesta descrição de dados foi que, aparentemente, os utilizadores recorrem à aplicação numa vertente mais social, já que a categoria mais visitada é *Food* e *NigthLife* - nos dias da semana/horas conforme a semana/hora vai avançando, existe um aumento de atividade.

4.3. Conclusão

Neste capítulo foi apresentado e estudado o conjunto de dados adotado. Foram ainda analisadas, descritivamente, as variáveis que serão utilizadas, tendo sido dado especial ênfase às variáveis de maior relevância, UserId e LocalID.

Importa destacar, também, os baixos valores encontrados no que respeita aos rácios de *check-in* por local e por utilizador.

5. Análise da Rede *Foursquare*

As métricas de análise de redes sociais podem ser baseadas na teoria de grafos permitindo analisar a estrutura e as relações da rede como um todo, subgrupos de atores e atores individualmente dentro da rede.

Uma rede pode ser analisada segundo dois âmbitos: a de rede, na qual a rede como um todo é analisada; individualmente, na qual cada nó é analisado.

O objetivo principal deste capítulo foi obter medidas referentes à rede *Foursquare*, tais como comprimento de caminhos mais curtos, graus de distribuição e grau máximo, oferecendo uma caracterização sobre a arquitetura da rede, além de identificar *hubs* e possíveis ligações entre locais.

Foram criadas redes de diferentes dimensões que permitissem um estudo detalhado da rede *Foursquare*: F10, F50, F100 e F500 (Tabela 5.1)

As redes formadas foram analisadas de acordo com os seguintes parâmetros:

Tabela 5.1 - Resultado dos parâmetros de cada rede

Locais	Utilizadores	Check-In	Check-In/Locais	Nome Rede
56.854	48.308	183.211	3,22	FS
500	15.980	26.671	53,34	FS500
100	8.087	10.732	107,32	FS100
50	5.809	7.209	144,18	FS50
10	2.424	2.792	279,2	FS10

Numa primeira análise à Tabela 5.1 nota-se que, a rede original (FS) contém um elevado número de locais e um rácio *check-in*/local muito baixo. Esse valor vai crescendo conforme a diminuição do número de locais, sendo que, a rede FS100, FS50 e FS10 apresentam já valores mais fáceis de serem analisados.

5.1. Caracterização da Rede

Com vista a identificar os nós mais relevantes na rede existiu a necessidade de recorrer a uma serie de estatísticas capazes de efetuar uma análise cuidada das redes. As medidas de redes são fundamentais para a sua caracterização, representação, classificação e análise.

Uma das medidas mais importantes é o grau médio da rede (*Average Degree*), que é calculado tendo em conta o Grau (*Degree*), o Grau Interno (*In-Degreee)* e o Grau Externo (*Out-Degree)*. O grau de um vértice mede o nível de envolvimento do mesmo na rede e determina o número de vizinhos do vértice. Neste caso teremos dois tipos de grau, o Grau Interno e o Grau Externo. No Grau Interno, a vizinhança é constituída por todos os vértices que estejam ligados ao vértice *v* (direcionados para v), enquanto que, para o Grau Externo, a vizinhança é definida por todos os vértices aos quais o vértice *v* se liga.

O grau médio ponderado dos nós (*Average Weighted Degree*) representa o número de ligações ponderadas que, em média, os nós de uma rede possuem. É uma medida que leva em consideração o peso/intensidade das ligações entre os nós. Logo, não avalia apenas se há conexão entre eles, mas sim a intensidade dessa ligação.

Em relação ao parâmetro Densidade de Rede, explica o nível geral de conectividade da rede e caracteriza-a como dispersa, quando possui uma densidade baixa. Representa a proporção de arestas (m) de um grafo (G) relativamente ao número máximo de arestas variando entre 0 e 1.

O Diâmetro da Rede permite calcular a maior distância existente entre dois nós nessa rede. Por distância, pode-se entender o número de ligações intermediárias existentes entre esses nós. Essa medida é útil para avaliar, em uma certa medida, a maior distância que os nós precisariam percorrer para se conectarem.

A medida HITS do software GEPHI analisa dois valores distintos para cada nó, o primeiro valor (denominado *Authority*) mede os nós que possuem popularidade e relevância. O segundo valor (denominado de *Hub*) quais nós são responsáveis por referenciar as autoridades e as unir.

A tabela 5.2 resume as medidas estatísticas utilizadas para este estudo e que irão ser interpretadas em cada uma das redes nos subcapítulos seguintes.

Tabela 5.2 - Resultado das medidas Referentes às Redes analisadas

Nome da Rede	Densidade	Grau Médio Ponderado	Diâmetro Rede	Comprimento Caminho Médio	*Caminhos curtos*
FS500	0,01	1.868,00	6	3,06	244.533
FS100	0,16	28,25	4	1,95	9.801
FS50	0,28	28,00	3	1,79	2.450
FS10	0,80	36,80	2	1,33	90

Após a realização das medidas, foram encontrados os dois *hubs* mais importantes para cada rede, ou seja, os locais que obtiveram mais ligações e de maior relevância na rede. Os locais, suas categorias e respetivo grau estão descritos na Tabela 5.3.

Tabela 5.3 - Os locais mais relevantes de cada rede

Rede	Local	Nome	Categoria	Grau
FS10	3	Aeroporto de Newark	*Travel Spots*	17
	5	One Pen Plazza	*Travel Spots*	17
FS50	5	One Pen Plazza	*Travel Spots*	52
	9	The Metropolitan Museum of Art	*Arts & Entertainment*	48
FS100	5	One Pen Plazza	*Travel Spots*	72
	1	Aeroporto John F. Kennedy	*Travel Spots*	65
FS500	9	The Metropolitan Museum of Art	*Arts& Entertainment*	74
	2	Aeroporto LaGuardia	*Travel Spots*	67

5.1.1. Rede FS10

A Figura 5.1 representa a rede FS10, mostrando uma estrutura generalizada, onde é possível identificar genericamente a presença de nós com mais visitas e com mais relevância na rede. Os maiores nós são os que tiveram maior destaque com os valores mais altos. As cores variam também em ordem crescente dos valores de grau de entrada, entre branco e roxo, ou seja, os menores valores são representados por cores mais claras enquanto que os maiores por cores mais escuras. O tamanho das setas que ligam os nós representam proporcionalmente o peso de cada aresta.

Os locais e as categorias representadas na rede são os seguintes*:*

Local *1 - Aeroporto John F. Kennedy (Travel* Spots)

Local 2 - O Aeroporto LaGuardia(*Travel Spots*)

Local 3 - Aeroporto Internacional de Newark(*Travel Spots*)

Local 4 - Yankee Stadium (*Arts & Entertainment*)

Local 5 - One Penn Plaza (*Travel Spots*)

Local 6 - O Grand Central Terminal (*Travel Spots*)

Local 7 - Museu de Arte Moderna (*Arts & Entertainment*),

Local 8 – Garden State Plaza *(Shops*)

Local 9 - The Metropolitan Museum of Art(*Arts & Entertainment*)

Local 10 - Union Square Park (*Great Outdoors*)

Analisando a rede FS10 onde o tamanho dos nós indica o valor do seu grau, claramente se consegue identificar os locais com maior número de ligações. Os nós que se encontram projetados no centro da rede são também os que contribuem em grande parte para o grau medio ponderado de 36,8. Na rede FS10 os nós pertencentes à categoria *Travel Spots* 1, 3 e 5, são os locais com maior número de ligações.

No cálculo da distância da rede é também calculado a distância média de um caminho, igual a 1,33, e o número de caminhos mais curtos existentes na rede, 90. Assim é possível concluir que em média é necessário passar por 1,33 nós para chegar de um determinado nó até outro. A análise do diâmetro da Rede FS10 que é de 2, poderá indicar que o número de passagens entre dois locais é bastante baixo.

As setas que ligam os nós avaliam se existem e que quantidade de visitas que um determinando local recebeu de todas os outros locais, calculando a soma dos pesos das arestas que outros nós possuem conectando-os uns aos outros. Neste caso, pode-se concluir que estamos perante uma rede densa, apresentando um valor de 0,8 de densidade, ou seja, quer isto dizer que em um quadro de 100% de relações possíveis (rede total), verificam-se 80% de interações entre os locais.

A medida denominada de HITS analisa dois valores distintos para cada nó, o primeiro valor (denominado *Authority*) mede quão valiosa é a informação armazenada no nó. O segundo valor (denominado de *Hub*) mede a qualidade das ligações de cada nó. Os nós com valores mais altos em ambas as medidas são, o nó 3 - Aeroporto Internacional de Newark (*Travel Spots*) e o nó 5 - One Penn Plaza (*Travel Spots*). Através da análise à Figura 5.2 podemos verificar que para além destes nós possuírem uma posição privilegiada são também os que possuem maior número de ligações a outros nós. A relação entre estes dois locais e os locais com mais ligações será detalhada no capítulo seguinte.

Por fim e de acordo com as estatísticas, pode-se afirmar que a rede tem uma estrutura densa e que em termos de arestas o número de passagens entre dois locais é baixo, dando assim origem a sequências muito curtas. Os nós com posição mais relevante e centrais são o nó 3 - Aeroporto Internacional de Newark (*Travel Spots*) e o nó 5 - One Penn Plaza (*Travel Spots)*.

5.1.2. Rede FS50

A Figura 5.2 representa a rede FS50. Nesta rede é possível visualizar e identificar nós predominantes, os que apresentam um número elevado de ligações. Maioritariamente esses nós, são os nós que constituem a rede FS10.

Analisando a Densidade da rede é possível identificar o quão perto a rede está de ser uma rede completa, ou seja com a totalidade dos nós ligados entre si.

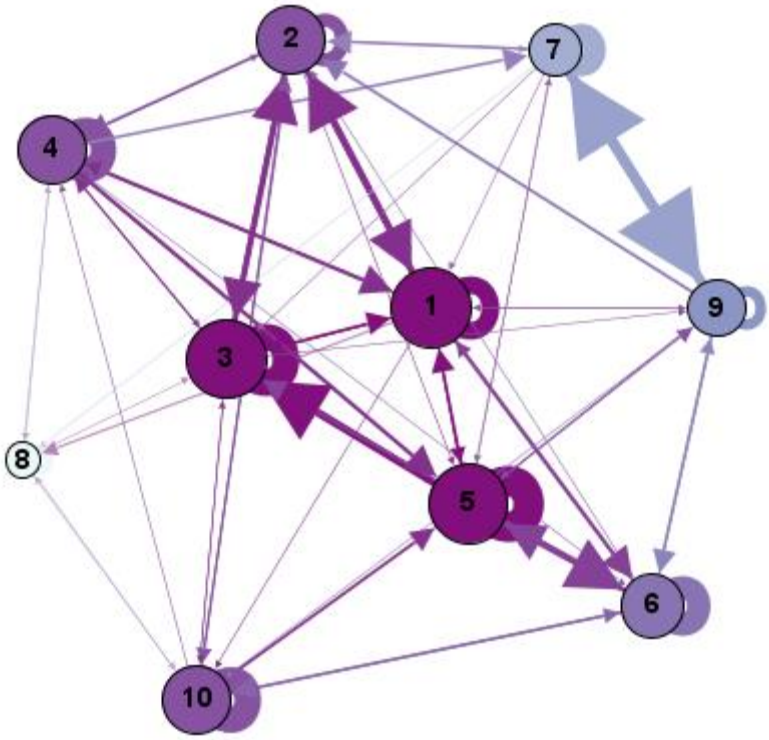

Figura 5.1 - Rede FS10

A densidade da rede FS50 é de 0,28 o que indica que nem metade das ligações são efetuadas, originando assim uma rede pouco densa.

O *Average Weighted Degree* da rede FS50 é de 28, onde os nós que contribuem mais significativamente para essa média são os nós 5,3,1 e 9.

A análise do diâmetro da rede fornece várias estatísticas relacionadas com as distâncias: o diâmetro, o comprimento do caminho médio e o número de caminhos mais curtos. Esta rede possui um diâmetro de 3, significando que a distância máxima entre qualquer par de nós é de apenas 3 nós.

Na ligação de qualquer par de nós, a distância média é de 1,79, ou seja, em média a distância das ligações entre quaisquer dois autores da rede é de 1,79 nós. A quantidade de caminhos mais curtos é de 2450.

Na rede FS50, os *hubs* que se caracterizam pela elevada quantidade de ligações que possuem, são eles o nó 1 - *Aeroporto John F. Kennedy (Travel* Spots),5 - One Penn Plaza (*Travel Spots*) e o nó 9 - The Metropolitan Museum of Art(*Arts & Entertainment*).

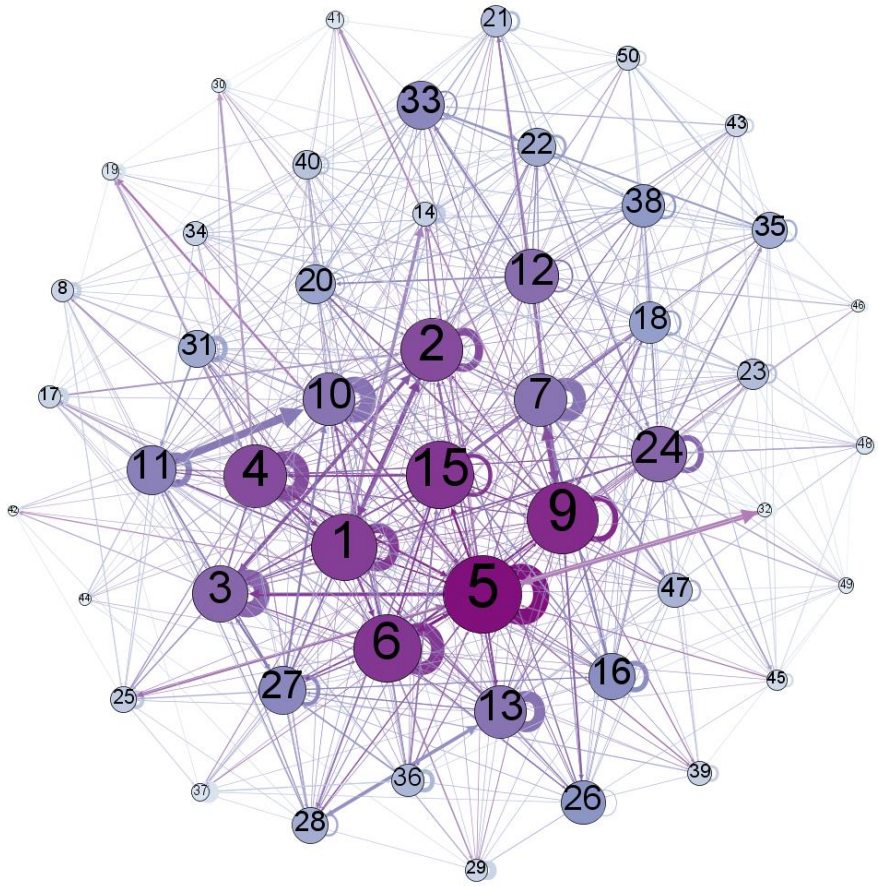

Figura 5.2 - Rede FS50

Para a medida *Authority*, foram identificados os mesmos locais na rede, os nós 1,5 e 9.

Verifica-se assim, que os dois nós com maior destaque na rede são dois nós pertencentes à rede FS10,ou seja, os nós 5 e 9 da categoria *Travel Spots* e *Arts & Entertainement.* A relação entre estes dois locais e os locais com mais ligações será detalhada no capítulo seguinte.

5.1.3. Rede FS100

A rede FS100, constituída por 100 nós está representada na Figura 5.3. O seu grau medio ponderado é de 28,25 e os nós mais importantes em termos de peso são os nós 1,5, e 9.

O Diâmetro da Rede FS100 é de 4, o que significa que a distância máxima entre qualquer par de nós é de apenas 4 nós. Para um total de 100 nós este valor não parece à primeira vista ser elevado e poderá ser justificado pela existência de pequenos grupos que se interligam entre si por meio de outros nós.

O comprimento médio do caminho, para a rede FS100 é de 1,955. No que respeita aos nós identificados como *Hubs* na rede FS100 pode-se mencionar os nós 5 - One Penn Plaza (*Travel Spots*) ,9 - The Metropolitan Museum of Art(*Arts & Entertainment*),1 - *Aeroporto John F. Kennedy (Travel* Spots) e 2 - O Aeroporto LaGuardia (*Travel Spots*). A relação entre estes dois locais e os locais com mais ligações será detalhada no capítulo seguinte.

À semelhança da rede FS50, os dois nós com maior destaque quer pelo seu peso quer pela sua relevância na rede são os nós 5 - One Penn Plaza (*Travel Spots*) e 1 - *Aeroporto John F. Kennedy (Travel* Spots). O nó 5 já referenciado na rede FS50 e o nó 1 tem sido um nó que tem vindo a obter bons resultados em todas as redes. Ambos pertencem à mesma categoria, *Travel Spots.*

5.1.4. Rede FS500

A Figura 5.4 representa a rede FS500, onde se podem identificar à primeira vista alguns nós com um grau de importância na rede. Os nós 9,5,1,22,12,2 e 23 contribuem de forma significativa para o cálculo do grau medio ponderado de 1868 da rede. A Densidade de Rede é de 0,019, valor inferior aos valores apresentados nas redes anteriores explicando o nível geral de conectividade da rede e caracterizando-a mais como dispersa do que como densa.

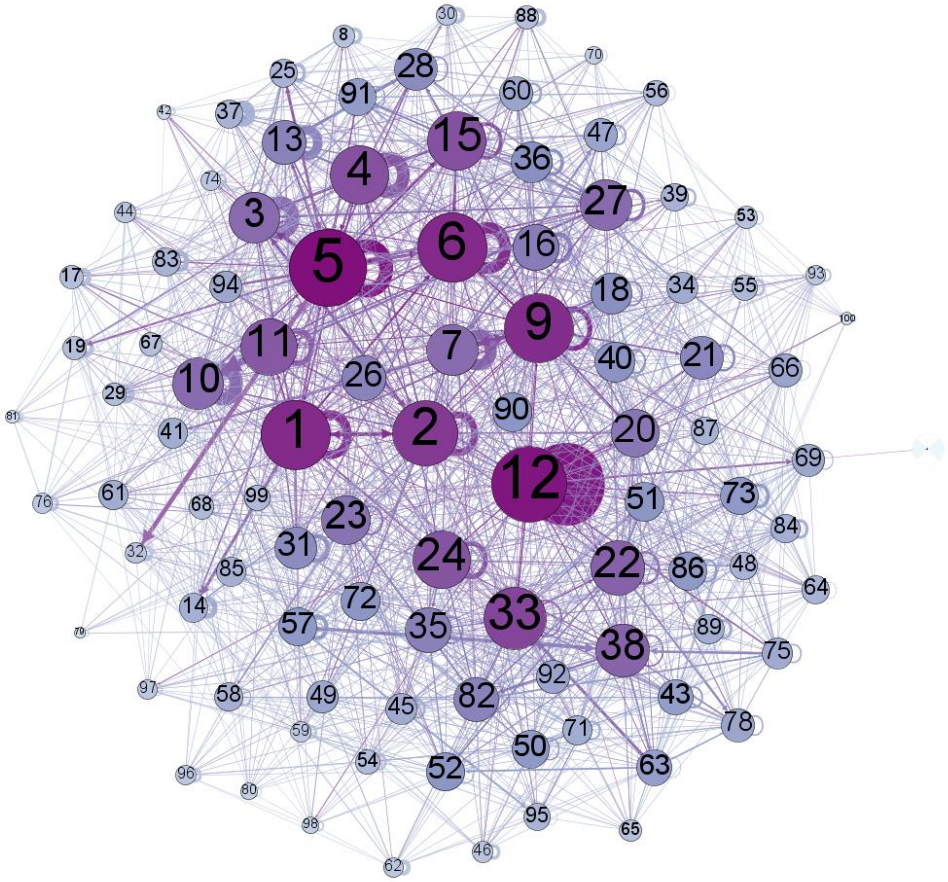

Figura 5.3 - Rede FS100

. O Diâmetro da Rede é de 6 que representa a maior distancia entre dois nós na rede e que parece ser um valor bastante aceitável na medida em que estamos a falar de uma rede bastante superior em termos de número de nós.

O comprimento de um caminho é definido pelo número de *links* entre dois nós. Na rede FS500 a media desse caminho é de 3,06 e o número de caminhos curtos na rede é de 244533.

Nas medidas *Authority* e *Hub* foram identificados os nós mais relevantes os nós 9,2,1,e 33. A relação entre estes dois locais e os locais com mais ligações será detalhada no capítulo seguinte.

Em relação à análise efetuada à rede FS500, existem alguns nós que se destacam com ótimos resultados como o nó 9 - The Metropolitan Museum of Art (*Arts &*

Entertainment) e o nó 2 - O Aeroporto LaGuardia(*Travel Spots*).Já era de esperar serem estes os locais mais relevantes dado que têm conseguido bons resultados em quase todas as análises efetuadas à rede.

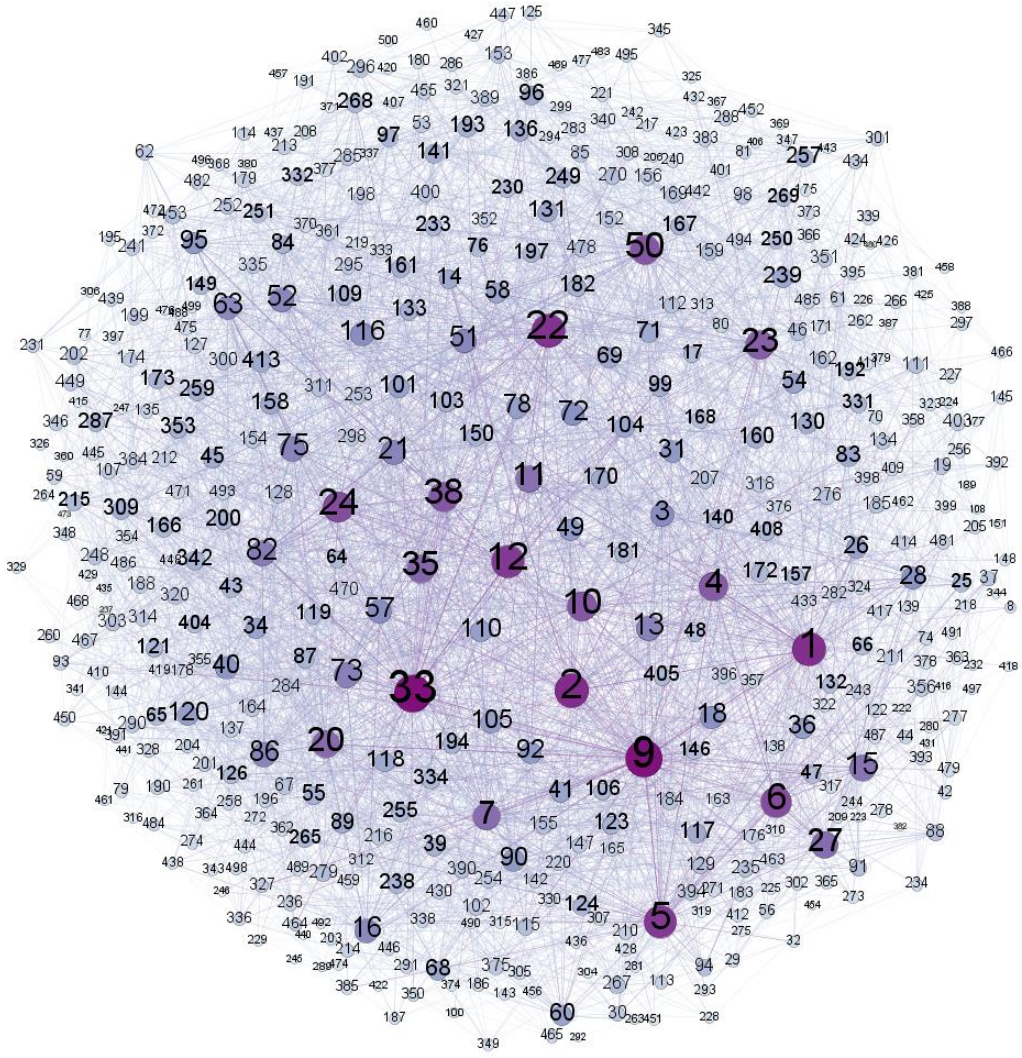

Figura 5.4 - Rede FS500

5.2. Visualização da rede utilizando *k-cores*

As estruturas de grafos muito grandes podem ser estudadas a partir da segmentação em partes menores, que são mais simples de serem manipuladas. Uma decomposição possível é baseada em k-*cores onde o* subgrafo induzido H_k é um *k-core*, ou núcleo de ordem k, se todo vértice de H_k possui grau menor ou igual à k e H_k for o maior grafo com esta propriedade. O *core* de maior ordem é chamado de *core* principal e o número de *core* de um vértice, v, é a menor ordem de core que contêm este vértice, denotado por $k_S(v)$.

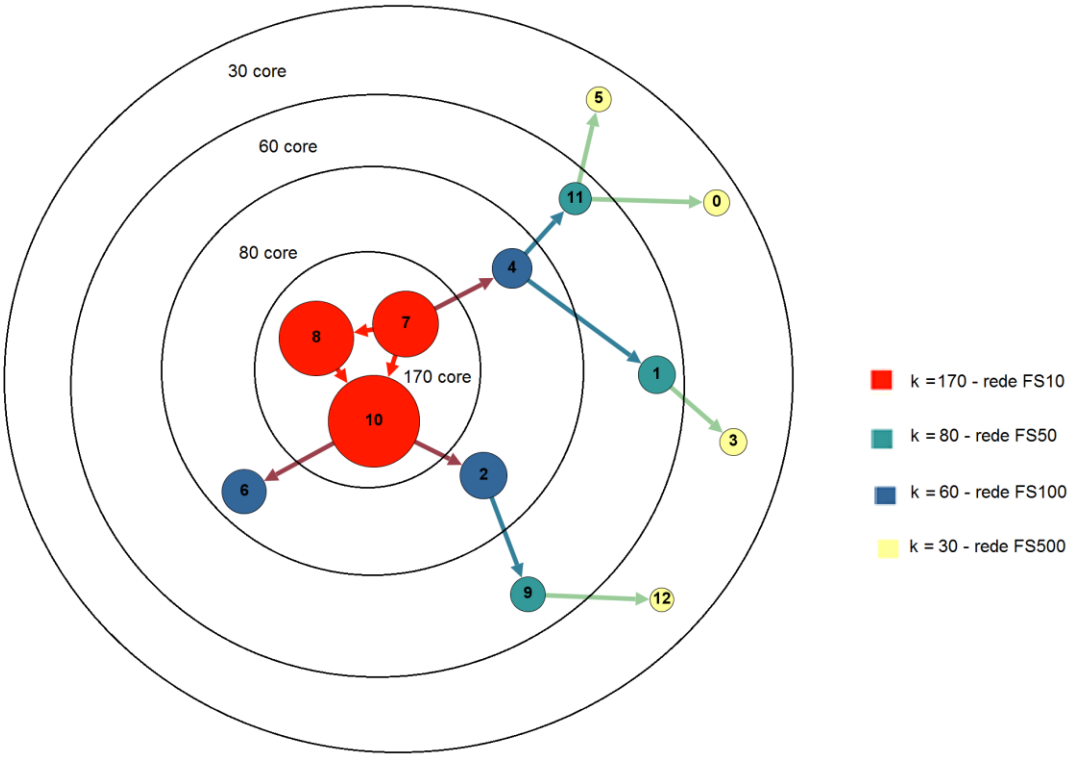

Figura 5.5 - Rede original com a divisão em núcleos

A metodologia utilizada nesta análise baseou-se em um estudo da rede que realiza a decomposição em *k-cores* baseado no número de *check-in*s. É baseado na remoção dos nós de acordo com o número de *check-ins,* onde k indica o número de *check-ins* mínimo que todo elemento do núcleo possui. Uma Rede *k-core* significa que todos os

locais têm visitas mínimas igual a k. A Figura 5.1 exemplifica o método de decomposição em núcleos.

Na primeira etapa exclui-se aqueles vértices com k inferior ou igual a 30. Os vértices que não foram removidos, bem como as arestas que conectam os mesmos formam o núcleo k_S = 30, ou seja, a rede FS500. A seguir, repete-se o mesmo procedimento para os demais locais presentes na rede para k inferior a 60, 80 e 170 respetivamente para as redes FS100, FS50 e FS10.

5.3. Conclusão

Neste capítulo foram criadas redes de diferentes dimensões que permitissem um estudo detalhado da rede. Foram identificados os nós com maior centralidade utilizando as métricas disponíveis no Gephi em particular as medidas Hit. Para obter uma visão mais geral foram utilizados os k-cores.

Das análises efetuadas às diferentes redes é verificável que para além de as redes serem pouco dispersas, existem alguns locais com relevância na rede. Se por um lado maior parte das redes apresentam uma baixa densidade por outro há entidades que nesta lógica de precariedade de relacionamentos já referidos, acabam mesmo assim por ocupar uma posição aparentemente privilegiada no quadro geral da rede.

Pelo número de caminhos curtos existentes em todas as redes pode-se afirmar que não existem sequências longas, os utilizadores dispersam-se pelos múltiplos locais e mesmo entre pares de locais não existem muitas passagens.

Apesar de encontrar os nós de maior centralidade a densidade gráfica das Figuras 5.1 a 5.4 não permitem identificar os percursos dos utilizadores com clareza. No capítulo seguinte, vamos utilizar o algoritmo Ramex para identificar a poli-árvores dos percursos dos utilizadores da rede *FourSquare*.

6. Análise dos Resultados

O presente capítulo tem como objetivo, apresentar uma análise dos resultados obtidos na fase de geração de poli-árvores pelo algoritmo Ramex [Cavique 2007, Cavique e Coelho 2008, Cavique 2015], aplicado às redes FS10, FS50, FS100 e FS500.

Para além de apresentar graficamente a análise das poli-árvores, este capítulo irá descrever o estudo das principais sequências encontradas na rede FS500.

A poli-árvore é um grafo orientado acíclico, com um arco entre cada par de nós no máximo. O grau interno dos vértices de uma árvore é zero (a raiz) ou um. Por sua vez, o grau interno dos vértices de uma poli-árvore pode ser maior do que um. Poli-árvores podem ser referenciadas como redes individualmente conectadas sendo possível a sua representação através de um grafo G = (V,A), formado por Vértices (V) e Arestas (A). Cada vértice ou nó representa um local e cada aresta representa a relação existente entre dois locais integrantes da rede.

Em cada rede estudada, estão identificados a laranja nós da rede superior, com o objetivo de identificar os nós comuns e o núcleo da própria rede.

6.1. Poli-árvore FS10

O algoritmo Ramex é composto por duas fases, a fase de transformação do *dataset Foursquare* numa rede cíclica e uma segunda fase de procura das sequências frequentes representadas numa poli-árvore. A figura 6.1 representa os locais mais frequentes na rede FS10, os dados de saída da segunda fase do Algoritmo Ramex aplicado a esta rede.

Como foi referido anteriormente, o *Graphviz* a ferramenta utilizada, permitiu a criação e visualização da poli-árvore, onde todos os elementos foram organizados de forma automática e os nós foram coloridos permitindo distinguir os diferentes tipos de categoria.

Foram observados algumas particularidades na poli-árvore, as quais poderão ser uteis na procura de padrões sequenciais e de descoberta de informação.

Na rede FS10 o nó 3 – Aeroporto de *Newark* e o nó 5- *One Penn Plaza* que foram considerados os locais com mais ligações, continuam a ter um lugar de destaque na poli-

árvore. Os locais com a mesma cor i.e. categoria, encontram-se agrupados e perto uns dos outros, o que se pode considerar ser um comportamento normal pelas razões já apresentadas como relações entre locais no capitulo anterior.

Como nó de origem podemos identificar, o nó 4 - Yankee Stadium e o nó 5- *One Penn Plaza* e como nós de destino o nó 2 - O Aeroporto LaGuardia, o nó 8 - Garden *State Plaza* , o nó 9 - *The Metropolitan Museum of Art* e o nó 10 - *Union Square Park*. Em relação aos nós de ligação o nó 1-*Aeroporto John F. Kennedy*, o nó 3 - Aeroporto Internacional de Newark,o nó 6 - O *Grand Central Terminal* e o nó 7- Museu de Arte Moderna, sendo o nó 2 - O Aeroporto *LaGuardia* o único nó de união.

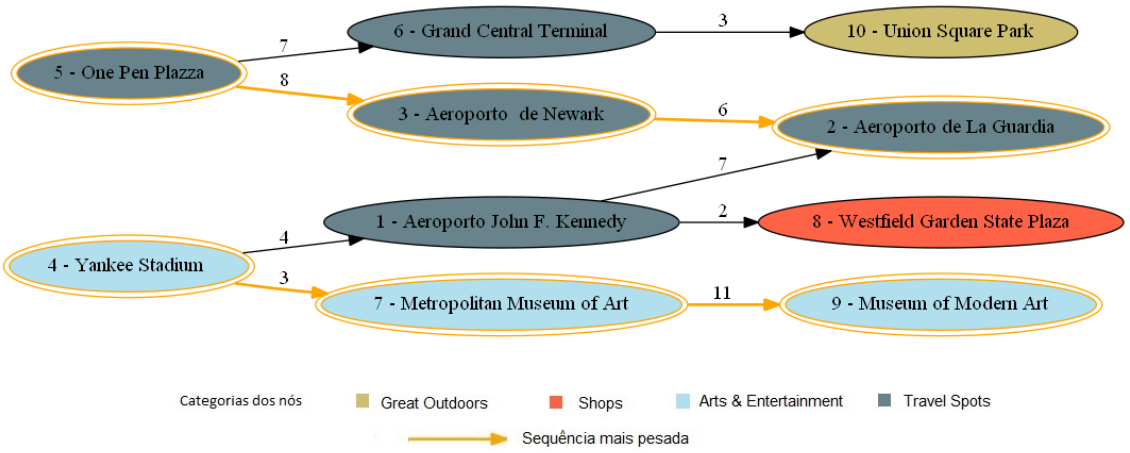

Figura 6.1 – Locais e Categorias da Rede FS10

Na poli-árvore resultante há a destacar a relação existente entre o local 7 -- The Museum of Modern Art e o local 9 - The Metropolitan Museum of Art que se localizam na mesma zona, junto ao Central Parque, a sua proximidade resulta no maior valor existente entre relação entre ambos os locais.

Importante mencionar que na rede FS10 as sequências mais pesadas são compostas pelos nós (4,7,9) e (5,3,2) com peso de 14, os quais estão devidamente assinalados a laranja. As sequências mais pesadas são constituídas apenas por 1 categoria, ou seja, todos os nós dentro dessa sequência pertencem à mesma categoria.

A árvore ponderada de maior peso (árvore que visita todos os nós) tem um comprimento total igual a 51 (8 + 7+ 6 + 7 + 4 + 3 + 11 + 3 + 2).

6.2. Poli-árvore FS50

Para a rede FS50, verificou-se através da análise da rede que os locais com mais destaque seriam os nós 5 e 9. Pela leitura da poli-árvore (Figura 6.2) verifica-se que ambos os nós são os locais com mais ligações a outros locais, estando localizados no centro da poli-árvore.

Na poli-árvore da rede FS50, consegue-se identificar os pontos da rede FS10 a laranja, mostrando que todos eles se encontram em posições de destaque e mostrando que esta rede continua a pertencer ao núcleo da rede FS50.

Os pesos das arestas não são muito elevados, mas as relações encontradas foram significativas mostrando que, mesmo para locais que não estão fortemente relacionados, ligações relevantes ainda são encontrados usando este algoritmo.

Embora os maiores pesos estejam diretamente relacionados com os nós pertencentes à rede FS10, os pesos entre os nós são distribuídos equitativamente, não se notando grandes diferenças de valores entre categorias e zonas da poli-árvore.

A relação entre locais mais significativa pertence ao local 11 - *Herald Square* e local 10 - *Union Square Park* que embora pertençam a categorias diferentes encontram-se muito próximos fisicamente. De notar ainda, uma sequência bastante completa (12,20,9,6,29), que abrange todas as categorias existentes na rede FS50 (*Food, Great Outdoors, Arts & Entertainment, Travel Spots, Home and Work*).

Tal como a poli-árvore resultante da rede FS10, os nós com categorias iguais tendem a estar próximos uns dos outros, dividindo a poli-árvore em dois grandes grupos. O primeiro grupo mais localizado no topo da poli-árvore, contém uma grande parte de locais da categoria *Food* e a zona mais abaixo as restantes categorias, *Travel Spots*, *Arts & Entertainment, Great Outdoors e Shops*.

Neste contexto, pode-se dizer que a zona superior por ser a zona com mais desmultiplicação de nós, pode ser a zona de utilizadores frequentes que usam a

aplicação de forma mais social. Enquanto que, a zona abaixo é a zona de utilizadores mais esporádicos da aplicação, ou seja, contém nós que originam sequencias menos profundas onde os utilizadores utilizam a aplicação de forma mais ocasional e numa vertente mais turística.

A categoria *Travel Spots* para além de ser a categoria com mais locais visitados pelo número de nós que compõem a poli-árvore e pelo peso que relaciona esses mesmos nós, é também a que tendencialmente surge como nó de origem.

A árvore ponderada de maior peso (árvore que visita todos os nós) tem um comprimento total igual a 211 para um total de 50 nós (locais) e 690 vértices (ligações entre locais).

6.3. Poli-árvore FS100

Na poli-árvore da rede FS100, representada na Figura 6.3 consegue-se identificar os pontos da rede FS50 a laranja, mostrando que todos eles se encontram em posições de destaque e mostrando que esta rede continua a pertencer ao núcleo da rede FS50.

Na rede FS100 onde os locais mais centrais e com mais peso são o nó 5 - *One Penn Plaza* e o nó 1 - *Aeroporto John F. Kennedy* verifica-se que, juntamente com o nó 3- Aeroporto Internacional de *Newark*, continuam a ser os nós que interligam e que acabam por serem comuns a vários caminhos ao longo de toda a poli-árvore.

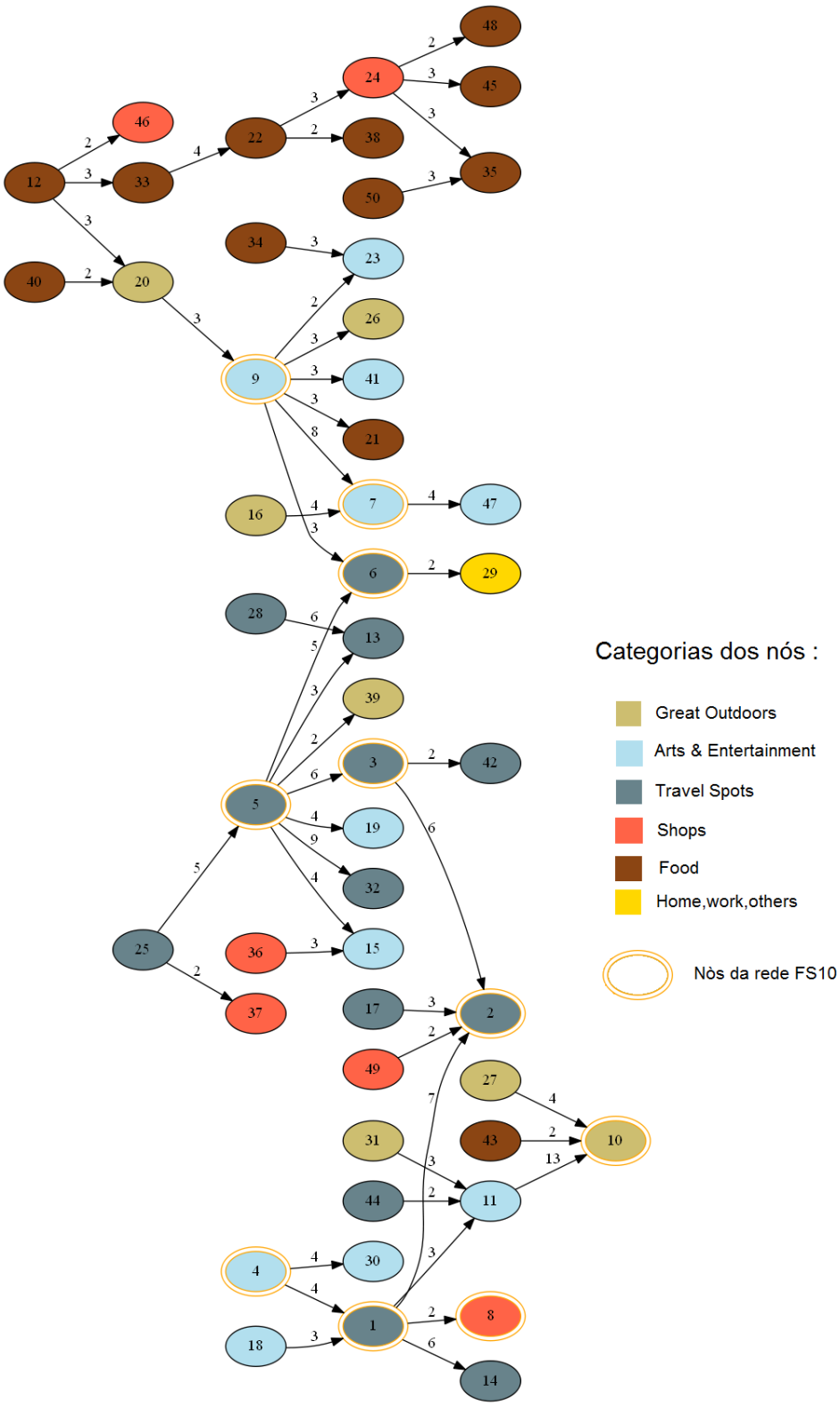

Figura 6.2 - Locais e Categorias da Rede FS50

Os pesos das arestas continuam sem apresentar valores muito elevados o que permite afirmar que não existem sequências com valores muito pesados, ou seja, que os utilizadores da aplicação dispersam-se um pouco entre a transição entre locais.

Mais uma vez se constata que os maiores pesos, embora não sejam valores muito relevantes em relação aos restantes, continuam de certa forma relacionados com nós pertencentes à rede FS10.

A categoria *Travel Spots* continua a ser a categoria com mais locais visitados pois para além de ser a categoria que tem maior representação é também onde se encontram a maioria dos nós com maiores pesos. No resultado da poli-árvore desta rede podemos assistir ao surgimento de novas categorias que até agora não tinham tido peso suficiente para surgir nesta estrutura, caso da *Nigthlife* Spots e *Home* and *Work*. Curiosamente surgem em posições esperadas, ou seja, como nós de origem ou de destino o que se considera um comportamento normal na medida em que são locais com características de início ou fim de percurso tal como já referenciado no capítulo das estatísticas do *dataset Foursquare*.

A tendência existente para a divisão da poli-árvore em dois grandes grupos continua a verificar-se. A predominância de ligações com mais profundidade na parte de cima da poli-árvore é onde os locais da categoria *Food* têm maior destaque.

Podíamos esperar que os utilizadores utilizassem racionalmente a rede *FourSquare* para identificar todos os pontos onde passam, gerando uma reportagem dos seus circuitos. Ora, esta presunção não existe na sua totalidade. Existem utilizadores que registam sequências de locais de categoria *Food* e outros que registam consecutivos locais de categoria *Travel*.

Uma das sequências identificadas é a (94,6,29) que reforça a ideia de que existem zonas da poli-árvore onde se conseguem identificar a natureza dos utilizadores, neste caso a sequência identifica um utilizador frequente que utiliza a aplicação para registar um percurso típico do dia-a-dia (*Food,Travel Spots, Home*). Um outro exemplo será a sequência (57,38,55) que representa um percurso entre locais de categoria *Travel, Food e Nightlife Spots.*

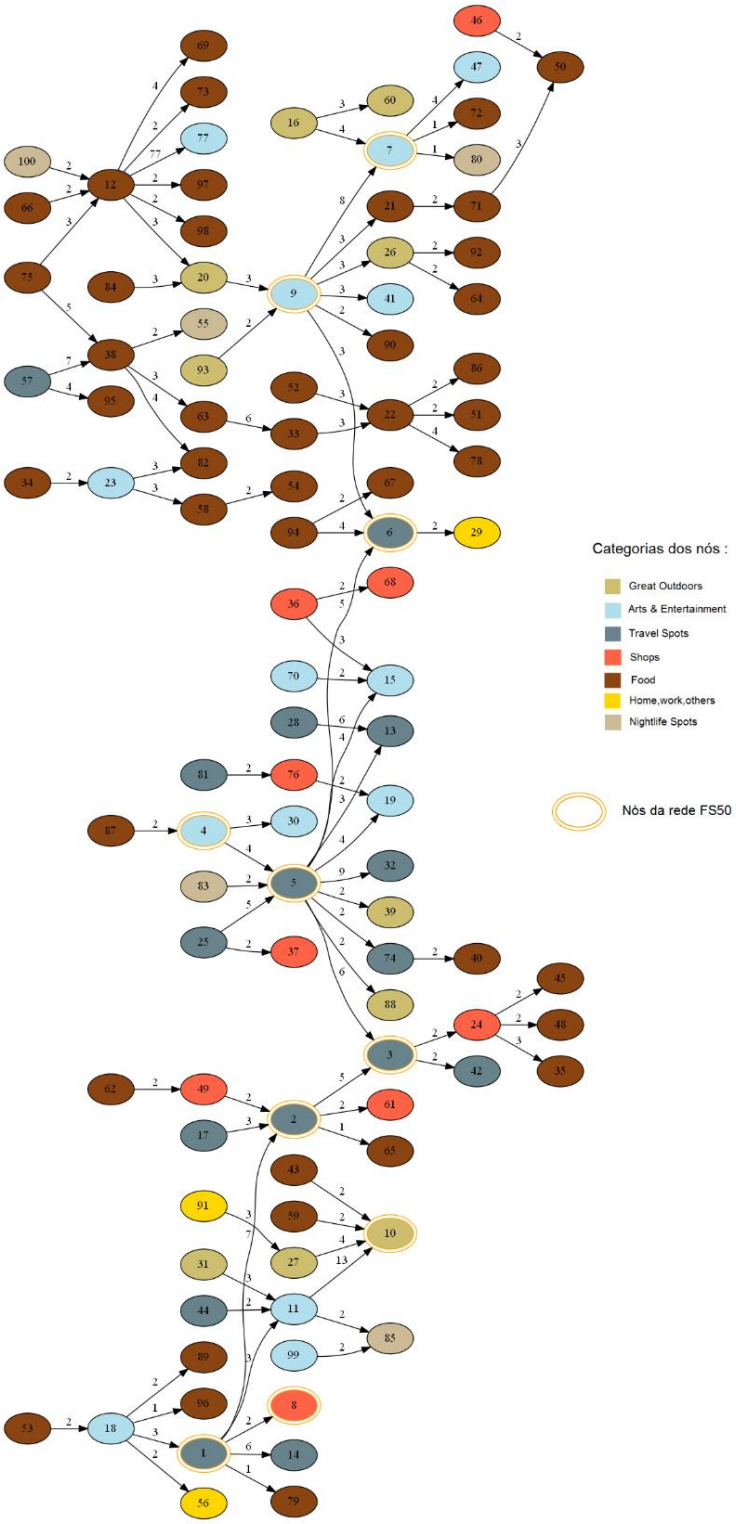

Figura 6.3 – Locais e Categorias da Rede FS100

A árvore ponderada de maior peso (árvore que visita todos os nós) tem um comprimento total igual a 454 para um total de 100 nós (locais) e 1662 vértices (ligações entre locais).

6.4. Análise das sequências de FS500

Devido à impossibilidade de apresentar graficamente o resultado da poli-árvore da rede FS500, optou-se por fazer um pequeno estudo das sequências geradas de forma a completar o estudo de todas as redes.

6.4.1. Analise das sequências por local

Pelo grande número de sequências geradas existiu a necessidade de definir um critério para filtrar as mais relevantes e significativas. Para contextualizar o problema pode-se dizer que, em uma base de dados de sequências, o objetivo será o de encontrar padrões sequenciais de acordo com o peso de uma aresta, onde é calculada uma percentagem relativamente à soma do peso de todas as arestas da poli-árvore.

As informações disponíveis a partir da poli-árvore são as seguintes:

- O número de locais visitados é de 500.
- O número máximo de visitas de um local a outro local é de 77.
- A média de número de visitas por local é de 2,2.
- A média de quilómetros percorridos entre dois locais é de 4,9.
- Soma do peso de todas as arestas da poli-árvore é de 1108.

Foram identificados os seguintes padrões de sequências:

- Uma sequência define que a visita ao local 12 - *Madison Square Park* é seguido de uma visita ao local 77- *World Trade Center*, com uma percentagem do peso da aresta de 6,9 %.

- Uma sequência define que a visita ao local 11 - *Herald Square* é seguido de uma visita ao local 10 - *Union Square Park* com uma percentagem do peso da aresta de 1,0 %.
- Uma sequência define que a visita ao local 5 - *One Penn Plaza* é seguido de uma visita ao local 32 – *Metro Station Sutphin Blvd - Archer Av* com uma percentagem do peso da aresta de 0,8 %.
- Uma sequência define que a visita ao local 7 - Museu de Arte Moderna é seguido de uma visita ao local 9 - *The Metropolitan Museum of Art* com uma percentagem do peso da aresta de 0,6 %.
- Uma sequência define que a visita ao local 2 - O Aeroporto *LaGuardia* é seguido de uma visita ao local 1 - *Aeroporto John F. Kennedy* com uma percentagem do peso da aresta de 0,5 %.
- Com uma percentagem do peso da aresta de 0,5 encontram-se as seguintes sequências: < {5, 3} >,< {10, 111} >,< {28, 13} >,< {57, 38} >,< {1, 14} >,< {2, 3} >,< {5, 6} >,< {10, 185} >,< {25, 5} >,< {481, 10} >.

6.4.2. Análise das sequências por categoria

Uma das grandes filosofias da aplicação *Foursquare* é a pesquisa detalhada sobre categorias de locais. Os locais estão sempre associados a categorias e em qualquer ação na aplicação está sempre implícita uma determinada categoria.

Baseado nesse facto, determinar o próximo local a visitar implica obrigatoriamente um estudo sobre as sequencias encontradas devidamente agrupadas por esta variável.

Com base na poli-árvore referente aos 500 locais mais visitados foram identificados a sua respetiva categoria. Na aplicação Microsoft Excel e com recurso a uma tabela dinâmica, foram agrupados e somado o peso dos diferentes nós por categoria.

Analisando a Figura 6.4 que representa o comportamento de transição entre categorias de locais, foram identificados os padrões gerais dos utilizadores da aplicação *Foursquare*.

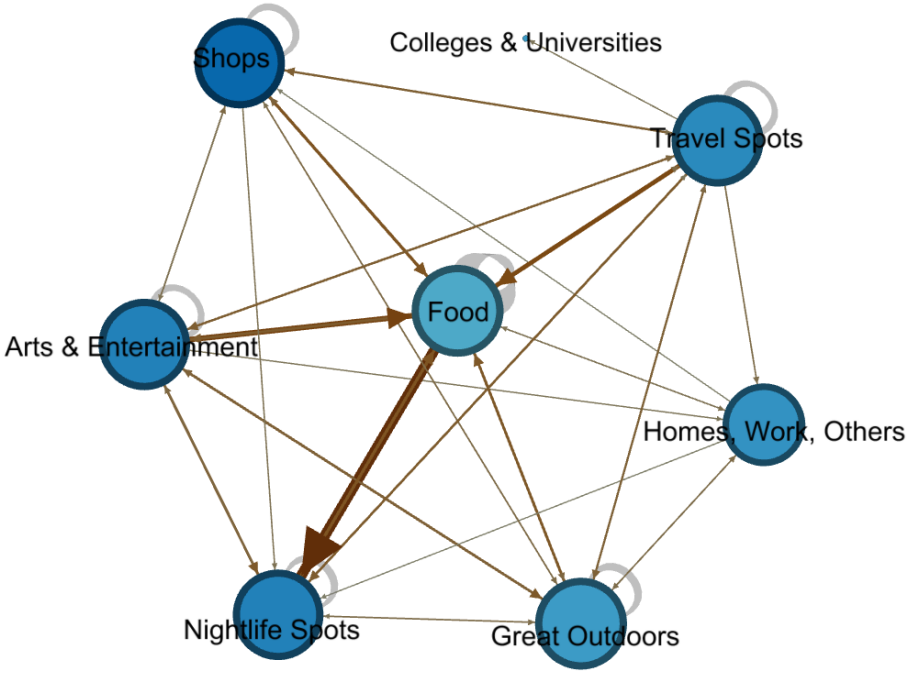

Figura 6.4 – Transição entre categorias da poli-árvore da rede FS500

As informações disponíveis são as seguintes:

- O número de categorias visitadas é de 8.
- O número máximo de visitas a uma categoria é de 539.
- A média de número de visitas por categoria é de 138.

Foram identificados os seguintes padrões:

- Uma sequência que define que, após uma visita a um local de categoria *Food* é seguido de uma visita a um local de categoria *Food* com uma percentagem do peso da aresta de 32,22%.
- Uma sequência que define que, após uma visita a um local de categoria *Food* é seguido de uma visita a um local de categoria *Nightlife Spots* com uma percentagem do peso da aresta de 11,91%.

- Uma sequência que define que, após uma visita a um local de categoria *Travel Spots* é seguido de uma visita a um local de categoria *Travel Spots* com uma percentagem do peso da aresta de 6,77%.
- Uma sequência que define que, após uma visita a um local de categoria *Arts & Entertainment* é seguido de uma visita a um local de categoria *Food* com uma percentagem do peso da aresta de 5,51%.
- Uma sequência que define que, após uma visita a um local de categoria *Arts & Entertainment* é seguido de uma visita a um local de categoria *Arts & Entertainment* com uma percentagem do peso da aresta de 4,69%.
- Uma sequência que define que, após uma visita a um local de categoria *Travel Spots* é seguido de uma visita a um local de categoria *Food* com uma percentagem do peso da aresta de 4,69%.

6.4.3. Análise por distância percorrida

Um dos filtros que se encontra disponível ao efetuar o *check-in* na aplicação é o da distância. Ter a possibilidade de configurar a distância a que se pretende efetuar uma visita em um determinado local é uma grande vantagem que a aplicação disponibiliza. A medição precisa entre dois locais visitados, pode trazer informações sobre o comportamento dos utilizadores da aplicação.

As informações disponíveis são os seguintes:

- O número total de quilómetros entre visitas aos locais é de 2460 quilómetros.
- O número máximo de quilómetros percorridos entre locais é de 39 quilómetros.
- A média de número de quilómetros por visita é de 4,92.

Através da leitura do gráfico da Figura 6.5 podemos verificar que a distância mais comum a ser percorrida entre dois locais é entre 2,5 e 4,99 quilómetros.

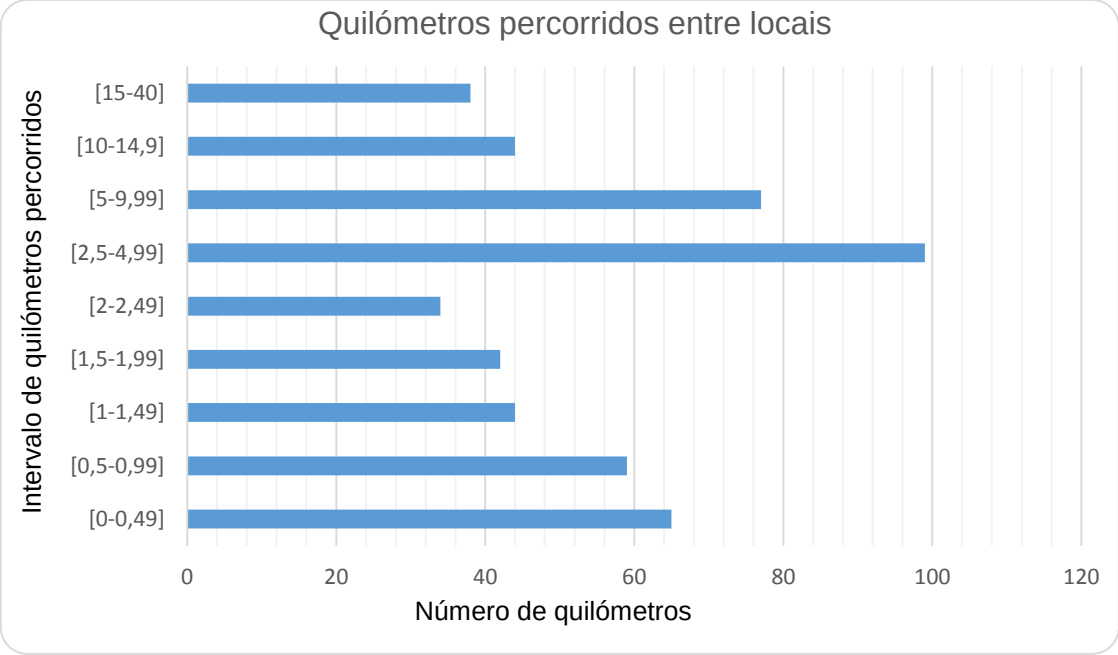

Figura 6.5 - Distribuição dos quilómetros percorridos

No entanto é importante enquadrar estas medidas com a análise efetuada no ponto anterior, ou seja, considerando as sequências por categoria do local. Tendo em conta isto, foram identificados os seguintes valores quanto às distâncias percorridas:

- Uma sequência que define que, após uma visita a um local de categoria *Food* é seguido de uma visita a um local de categoria *Food*, a distância percorrida em média é de 2,63 quilómetros.
- Uma sequência que define que, após uma visita a um local de categoria *Food* é seguido de uma visita a um local de categoria *Nightlife Spots* a distancia percorrida em média é de 2,49 quilómetros.
- Uma sequência que define que, após uma visita a um local de categoria *Travel Spots* é seguido de uma visita a um local de categoria *Travel Spots* a distancia percorrida em média é de 6,53 quilómetros.
- Uma sequência que define que, após uma visita a um local de categoria Arts & *Entertainment* é seguido de uma visita a um local de categoria *Food* a distancia percorrida em média é de 6,45 quilómetros.

- Uma sequência que define que, após uma visita a um local de categoria *Arts & Entertainment* é seguido de uma visita a um local de categoria *Arts & Entertainment* a distancia percorrida em média é de 5,69 quilómetros.
- Uma sequência que define que, após uma visita a um local de categoria *Travel Spots* é seguido de uma visita a um local de categoria *Food Spots* a distancia percorrida em média é de 11,60 quilómetros.

Verificou-se que, tendo em conta as categorias dos locais visitados os valores em média percorridos entre locais saem um pouco do intervalo de [2,5-4,99] quilómetros, dos 500 locais visitados.

6.5. Conclusão

Como conclusão deste capítulo, e baseado na análise visual e estatística das redes estudadas anteriormente, podemos dizer que estamos perante uma rede onde os utilizadores procuram os seus próximos locais de interesse e que se dispersam pelos diferentes locais na rede criando percursos bastante distintos conforme a frequência com que utilizam a aplicação.

As poli-árvores-soluções das redes FS50 e FS100 encontradas estão muito próximas da poli-árvore FS10. Todas as poli-árvores possuem características idênticas e todas elas têm em comum possuírem arestas entre os nós com valores não muito elevados e possuírem duas grandes zonas onde algumas categorias se destacam. A profundidade da desmultiplicação dos nós faz com que se consiga detetar dois tipos de utilizadores onde a frequência de *check-ins* difere conforme a intenção de utilização da aplicação.

Os resultados obtidos através da análise das poli-árvores foram coerentes com o que se esperava, onde se verifica que existem um conjunto de nós com grande importância nos curtos caminhos percorridos.

Nota-se de certa forma que as arestas que possuem maior grau de possibilidade de participar na árvore geradora de maior peso são aquelas que estão presentes nas melhores árvores encontradas, provando a coerência dos resultados obtidos.

Quanto às sequências detetadas, verificou-se que todas elas apresentam uma percentagem relativa bastante baixa o que leva a concluir que cada utilizador procura os seus próximos locais de interesse, afastando-se dos restantes.

Quanto ao comportamento dos utilizadores agrupados por categoria, consegue-se perceber uma clara tendência para visitas entre locais da categoria *(Food -> Food)* o que reforça a ideia de que esses utilizadores utilizam a aplicação diariamente e de forma mais social, partilhando para os seus contactos a sua localização e as suas preferências.

Seria expetável que os utilizadores da rede *Foursquare* fizessem uso da aplicação da sua totalidade, no entanto tal situação não se verifica. Através da análise detetou-se que existem utilizadores a usarem a aplicação para registar a sua atividade em duas grandes vertentes: registo de sequências em locais da categoria *Food* e locais relacionados com a categoria *Travel*.

7. Conclusões

Esta dissertação de mestrado assumiu como objetivo a aplicação de técnicas de *Data e Graph Mining* em Redes Sociais, com o intuito de permitir o estudo dos próximos Locais a Visitar na Rede *Foursquare*.

Uma das principais inquietações iniciais deste estudo foi sem dúvida o grande volume dos dados que o *dataset* apresentava.

Após inúmeras experiências com os dados pré-processados, verificou-se que a divisão do *dataset* em núcleos mais pequenos originando redes de diferentes dimensões iria contribuir de forma significativa para o objetivo deste estudo. Sobre estas diferentes redes foi aplicado o algoritmo Ramex [Cavique 2007, Cavique e Coelho 2008, Cavique 2015] e os resultados obtidos levam a acreditar no sucesso da aplicabilidade do modelo em situações práticas que requerem o estudo do comportamento humano.

Este trabalho apresentou uma nova abordagem ao uso de algoritmos que permitam uma visão global dos dados de uma rede possibilitando a visualização dos dados organizados numa poli-árvore de forma a obter sequências frequentes.

As técnicas de *Data Mining* foram aplicadas através de um estudo de caso único utilizando um *dataset* da aplicação *Foursquare*.

A etapa de pré-processamento dos dados que incluiu a limpeza e tratamento dos dados foi feita com recurso ao Software SAS [SAS, 2014]. Durante esta fase foram efetuadas operações de exclusões de registos impossibilitava a análise gráfica e o estudo de padrões das sequências, originando uma redução bastante significativa no volume de dados do *dataset*. A fase de compreensão dos dados permitiu a identificação da informação relevante contribuindo de forma bastante positiva para a fase de análise das medidas descritivas do *dataset*. É ainda de realçar, e ainda recorrendo às potencialidades do SAS, a decomposição do *dataset* em vários núcleos de forma a obter um estudo mais detalhado das várias redes a estudar.

As estruturas resultantes dessa decomposição foram estudadas a partir do Software Gephi [Gephi, 2015] que permitiu obter as medidas de análise das diferentes redes e identificar os locais mais destacados de cada rede. Baseado nessa análise visual e estatística, concluiu-se que existe uma grande dispersão dos percursos dos utilizadores, fazendo com que não existam longas sequências e que a transição entre

locais seja baixa. Os locais mais relevantes pertencem à categoria *Travel Spots* e são comuns a maior parte das redes estudadas.

Os resultados obtidos na fase de geração de poli-árvores pelo algoritmo Ramex apontam sobretudo para uma rede onde os utilizadores procuram os seus próximos locais de interesse e que se dispersam pelos diferentes locais na rede criando percursos bastante distintos conforme a frequência com que utilizam a aplicação.

A aplicação do Algoritmo Ramex à rede *Foursquare* obteve bons resultados essencialmente no que respeita à descoberta de sequências frequentes na poli-árvore de maior peso, objetivo subjacente a este trabalho. As sequências descobertas apontam para uma certa tendência onde os utilizadores procuram os seus próximos locais de interesse afastando-se dos restantes, criando diferentes percursos.

Embora as sequências descobertas não tenham resultado em valores absolutos muito altos, as ligações entre locais foram devidamente analisadas apresentando possíveis justificações para a sua relação.

Detetou-se também através da análise, que os utilizadores usam a aplicação para registar a sua atividade em duas grandes perspetivas: a perspetiva de utilizador frequente que regista sequências em locais da categoria *Food* e a perspetiva de utilizador ocasional que efetua *check-ins* em locais relacionados com a categoria *Travel*. Desta forma, exclui-se a hipótese de que os utilizadores da rede *Foursquare* utilizem a aplicação na sua totalidade e que cada utilizador cria os seus próprios percursos em determinadas categorias.

Apesar da satisfação global com os resultados obtidos, uma das melhorias sugeridas em trabalhos futuros, diz respeito à possibilidade de gerar a Árvore Geradora Mínima para a rede *Foursquare*, de forma a criar um registo dos locais menos visitados da rede. Uma área de interesse para empresas de estudos de mercado e/ou para campanhas de Marketing empresarial de forma a inverter uma tendência menos favorável em visitas de um determinado local.

8. Bibliografia

[Agrawal & Srikant, 1995] R.Agrawal, R.Srikant - **Mining Sequential Patterns.** Proceedings 11th International Conference Data Engineering, ICDE, pp. 3–14, IEEE Press, 1995.

[Bao, Zheng & Mokbel, 2012] J. Bao, Y. Zheng, M. Mokbel - **Location-based and Preference-Aware Recommendation Using Sparse Geo-Social Networking Data**. In ACM SIGSPATIAL (GIS 2012), pp.199-208, Redondo Beach, CA, US, 2012.

[Cavique, 2007] L. Cavique - **A network algorithm to discover sequential patterns**. In Proceedings of the Artificial intelligence 13th Portuguese conference on Progress in artificial intelligence, EPIA'07, LNAI 4874, Springer-Verlag Berlin Heidelberg, pp. 406-414. 2007.

[Cavique & Coelho, 2008] L.Cavique, J.Coelho - **Descoberta de Padrões Sequenciais utilizando Arvores Orientadas**. Revista de Ciências da Computação, volume III, ano III, nº3, pp. 12-22, 2008.

[Cavique, 2015] L.Cavique - **Ramex: A Sequence Mining Algorithm Using Poly-trees.** New Contributions in Information Systems and Technologies, Advances in Intelligent Systems and Computing, A.M Rocha, S. Correia, L.P Costanzo, Eds Reis, Springer edition, volume 2, pp. 143-154, 2015.

[Cheng et al., 2011] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui - **Exploring Millions of Footprints in Location Sharing Services**. In Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM, pp. 81–88, 2011.

[Edmonds, 1967] J.Edmonds - **Optimum Branchings.** J. Research of the National Bureau of Standards 71B, pp. 233- 240, 1967.

[Foursquare, 2014] Blog do Foursquare, http://blog.foursquare.com/ [Setembro de 2014]

[Fayyad et al., 1996] U.Fayyad, G.Piatetsky-Shapiro, P. Smyth, R. Uthurusamy - **Advances in Knowledge Discovery and Data Mining.** American Association for Artificial Intelligence Menlo Park, USA, 1996.

[Fulkerson, 1974] D.R Fulkerson - **Packing Rooted Directed cuts in a Weighted Directed Graph**. Mathematical Programming 6, pp.1-13, 1974.

[Gephi, 2015] Página oficial Gephi http://gephi.github.io/ [Fevereiro 2015]

[Graphviz, 2015] Página oficial Graphviz http://www.graphviz.org/ [Fevereiro 2015]

[Knuth, Morris &. Pratt, 1977] D. Knuth, J. Morris, V. Pratt - **Fast Pattern Matching in Strings**. SIAM Journal of Computing vol 6, pp. 323-350, 1977.

[Lin, Orgun & Williams, 2002] W. Lin, M. Orgun, and G. Williams - **An overview of Temporal Data Mining**. In Proceedings of the 1st Australian Data Mining Workshop, pp. 83-90, University of Technology, Sydney, 2002.

[Mannila, Toivonen & Verkamo, 1997] H. Mannila, H. Toivonen, A. Verkamo - **Discovery of frequent episodes in event sequences.** Data Mining and Knowledge Discovery, vol. 1, Issue 3, pp. 259-289, 1997.

[Marques & Cavique, 2013] N.C Marques, L.Cavique - **Sequential pattern mining of price interactions**, in EPIA 2013, 16th Portuguese Conference, Advances in Artificial Intelligence, Local Proceedings, Angra do Heroísmo, Açores, Portugal, 2013, pp. 314-325, 2013

[Milgram, 1967] S. Milgram. **The small world problem**. Psychology Today, vol. 2, pp. 60-67, 1967.

[Mascolo, et al., 2011] C. Mascolo, A. Noulas, S. Scellato, M. Pontil - **An empirical study of geographic user activity patterns in foursquare.** In International Conference on Weblogs and Social Media, Barcelona, pp.570-573, 2011.

[Moreno, 1934] J. Moreno - **Who Shall Survive?.** Beacon House, Beacon, NY, 1934.

[Noulas et al., 2011] A. Noulas, C. Mascolo, S. Scellato, and M. Pontil - **Exploiting semantic annotations for clustering geographic areas and users in location-based social networks.** ICWSM International Workshop on Social Mobile Web,SMW, pp. 570-573, 2011.

[Pietro & Cohn, 2013] D. Pietro, T.Cohn - **Mining user behaviours: A study of check-in patterns in location based social.** In International AAAI Conference of WebSci'13, pp. 306-315, 2013.

[Rapoport, 1957] A. Rapoport - **Contribution to the theory of random and biased nets**. Bulletin of Mathematical Biophysics, vol. 19, pp. 257-277, 1957.

[Sadilek & Krumm, 2012] A. Sadilek, J. Krumm - **Far out: predicting long-term human mobility**. Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 814-820, 2012.

[SAS, 2014] Página oficial SAS, http://support.sas.com/documentation/, [Julho de 2014].

[Srikant &.Agrawal, 1995] R. Srikant, R. Agrawal - **Mining sequential patterns: Generalizations and performance improvements**. Proceedings 5th International Conference Extending Database Technology, EDBT, 1057, pp. 3-17, 1996.

[Scellato et al., 2010] S. Scellato, C. Mascolo, M. Musolesi, V. Latora - **Distance Matters: Geo-social Metrics for Online Social Networks**. Proc. of the International Conference on Online Social Network,WOSN'10, pp.8-8, 2010.

[Sutko & Silva, 2011] D. Sutko, A. de Souza E Silva - **Location-aware mobile media and urban sociability**. New Media & Society, v. 13, pp. 807-823, 2011.

[Tiple, 2014] P.Tiple - **Tool for Discovering Sequential Patterns in Financial Markets**. Dissertação para obtenção do Grau de Mestre em Engenharia Informática, na Faculdade de Ciências e Tecnologia da Universidade Nova Lisboa, 2014.

[Tiple, Cavique & Marques, 2015] P. Tiple, L. Cavique, N. Marques - **Ramex-Forum: Sequential Patterns of Prices in the Petroleum Production Chain**. EPIA 2015, 17th Portuguese Conference, Advances in Artificial Intelligence, (accepted 2015).

[Vasconcelos et al., 2012] M. Vasconcelos, S. Ricci, J. Almeida, F.Benevenuto, V. Almeida - **Caracterização e influência do uso de tips e dones no foursquare**. Simposio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC), pp. 478-491, 2012.

[Wikipedia, 2014] Página oficial Wikipédia, http://pt.wikipedia.org, [Julho 2014].

[Wikipedia, 2015] Página oficial Wikipédia, http://pt.wikipedia.org, [Março 2015].

[Zago, Recuero, 2011] G. Zago, R. Recuero - **Usos e Apropriações do Foursquare no Brasil**. **Apontamentos para Discussão.** Revista Nexi, número 1, volume 1, 2011.

[Zaki, 2001] M.Zaki - **Spade: An efficient algorithm for mining frequent sequences**, Machine Learning, vol. 42, pp. 31–60, 2001.

[Zarur, 2005] M. Zarur - **Modelo para Elaboração de Cenários do Setor Energético, Utilizando Técnicas de Data Mining**. Dissertação de Mestrado (UFRJ), 2005.

[Zichermann & Linder, 2010] G.Zichermann, J.Linder - **Game-Based Marketing: Inspire Customer Loyalty Through Rewards, Challenges, and Contests**. Wiley, Hoboken, NJ, 2010