UNIVERSIDADE FEDERAL DO MARANHÃO

CENTRO DE CIENCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE ELETRICIDADE

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

RODRIGO MIRANDA FEITOSA

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS PARA RECOMENDAÇÃO SOCIAL

São Luís
2013

RODRIGO MIRANDA FEITOSA

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS PARA RECOMENDAÇÃO SOCIAL

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Eletricidade, área de concentração: Ciência da Computação.

Orientador: Prof. Dr. Sofiane Labidi

São Luís
2013

Feitosa, Rodrigo Miranda.

Uma aplicação de mineração de dados para recomendação social/ Rodrigo Miranda Feitosa – São Luís, 2013.

154 f.

Impresso por computador (fotocópia).

Orientador: Sofiane Labidi.

Dissertação (Mestrado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2013.

1. Mineração de dados. 2. Sistemas de recomendação. I. Título.

CDU 004.8

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS PARA RECOMENDAÇÃO SOCIAL

Rodrigo Miranda Feitosa

Dissertação aprovada em 22 de março de 2013.

Prof. Sofiane Labidi, Dr.
(Orientador)

Prof. Pedro Porfírio Muniz Farias, Dr.
(Membro da Banca Examinadora)

Prof. Zair Abdelouahab, Ph.D.
(Membro da Banca Examinadora)

Aos meus pais Raimundo e Raimunda.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus por todas as bênçãos derramadas sobre mim desde o início do percurso do Mestrado e que permitiram o desenvolvimento desta Dissertação.

Aos meus Pais por toda dedicação e zelo, para que a educação sempre fosse um fator primordial na minha vida e do meu irmão. E ainda, pela compreensão nos momentos de ausência nestes anos.

Ao meu orientador Professor Sofiane Labidi que me acolheu no Laboratório de Sistemas Inteligentes e garantiu esta grande oportunidade em minha vida. Obrigado pela paciência, dedicação e confiança a mim depositada.

Ao professor André Santos que me co-orientou nesta pesquisa e antes mesmo do Mestrado sempre me aconselhou e orientou em algumas perspectivas acadêmicas, obrigado pela sua amizade e bondade.

A minha namorada Valéria Raquel pelas palavras de encorajamento e carinho que me fortaleceram durante esta caminhada de estudo.

Ao pesquisador Marcel Caraciolo que me elucidou sobre o tema da Dissertação e me apontou um roteiro para a pesquisa.

Aos amigos do Mestrado e do Laboratório de Sistemas Inteligentes que participaram junto comigo dos momentos de dificuldade e alegria.

Aos companheiros Andrei Cunha, Ercília Maria e Keyne Conceição pela ajuda e colaboração no momento que precisei.

A todos os professores do curso de Mestrado que não mediram esforços em compartilhar com seus alunos o conhecimento adquirido.

A todos aqueles que me ajudaram de forma direta ou indiretamente na realização deste trabalho de pesquisa e conclusão do curso.

“Todos estamos matriculados na escola da vida, onde o mestre é o tempo”

(Cora Coralina).

RESUMO

A busca do conhecimento e a sua manipulação em empresas, instituições ou outras organizações tem se tornado um desafio nos dias atuais. Em grande parte devido a dois aspectos: o grande volume de informação disponibilizada e a dificuldade em extrair o conhecimento próprio de cada pessoa (capital intelectual). Essa dificuldade torna-se mais acentuada quando o cenário envolvido para a extração de conhecimento é a *Web.* A área da Gestão de Conhecimento busca a solução para as limitações descritas anteriormente. Técnicas para a extração e controle do conhecimento podem ser adotadas com o uso da Inteligência Artificial, sobretudo a Descoberta de Conhecimento em Bases de Dados.

Este trabalho propõe-se a criação de uma metodologia e aplicação que realize a Mineração de Dados com informações textuais vinculados a dados geolocalizados em uma Rede Social, com o intuito de promover a Recomendação Social. Entretanto, as abordagens na construção dos Sistemas de Recomendação apresentam algumas deficiências na filtragem dos resultados e na forma que estes são sugeridos aos usuários. A pesquisa busca a solução destas deficiências e aborda temas que ainda carecem de pesquisas mais efetivas e resultados consolidados.

Palavras-Chaves: Mineração de Dados, Sistemas de Recomendação, Rede Social Baseada em Localização e Recomendação Social.

ABSTRACT

The search of knowledge and its manipulation in companies, institutions or other organizations has become a challenge nowadays. Mostly due to two aspects: the large volume of information available and the difficulty in extracting the knowledge proper to each person (intellectual capital). This difficulty becomes more accentuated when the scenario involved the extraction of knowledge is the Web. The area of Knowledge Management seeks a solution to the limitations described above. Techniques for extracting and control of knowledge can be adopted with the use of Artificial Intelligence, particularly the Knowledge Discovery in Databases.

This work proposes the creation of a methodology and application that perform the Data Mining with textual information linked to geo data in a social network, in order to promote Social Recommendation. However, approaches in building recommendation systems present some shortcomings in filtering the results and the way they are suggested to users. The research aims to remedy these deficiencies and addresses issues that still need to search more effective and consolidated results.

Keywords: Data Mining. Recommender Systems. Location-Based Social Networking and Social Recommendation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

API	*Application Programming Interface*
BSD	*Berkeley Software Distribution*
GPL	*General Public License*
GPS	*Global Positioning System*
HITS	*Hyperlink-Induced Topic Search*
HTML	*HyperText Markup Language*
HYTASO	*Hybrid Tagging Social*.
IDF	*Inverse-Document-Frequency*
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Databases*
KDT	*Knowledge Discovery in Texts*
LBSN	*Location-Based Social Network*
LSI	Laboratório de Sistemas Inteligentes
NLTK	*Natural Language Toolkit*
PSF	*Python Software Foundation*
TF	*Term-Frequency*
TF-IDF	*Term-Frequency- Inverse-Document-Frequency*
UFMA	Universidade Federal do Maranhão
UML	Unified Modelling Language
XML	*eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

Este Trabalho de Dissertação está alinhando a área de Gestão de Conhecimento tendo como uma de suas técnicas a Mineração de Dados. A pesquisa ainda contempla o estudo de outras áreas pertinentes como: Redes Sociais e Sistemas de Recomendação.

1.1 Problemática

O conhecimento se apresenta em uma quantidade numerosa de dados, em cenários heterogêneos (Navegação na internet, Catálogos online, Ferramentas de busca, Redes Sociais e etc.). A Gestão do Conhecimento é um campo que estuda a forma que as organizações gerenciam aquilo que elas conhecem, o que necessitam ainda conhecer e como irão utilizar este conhecimento. É um processo amplo e nas últimas décadas alguns pesquisadores têm adotado o uso de técnicas de Inteligência Artificial para a Gestão de Conhecimento, de forma extensiva, sobretudo em alguns tópicos como: sistemas especialistas, modelo de conhecimento, integração de documentos e sistemas inter-organizacionais (Labidi, 1997).

A Mineração de Dados se apresenta como forma de garantir a descoberta de padrões que interessam na busca dos dados em questão. O papel da Mineração de Dados, no que diz respeito ao tratamento do conhecimento, é aplicar algoritmos sobre os dados e usando da abstração gerar modelos de conhecimento através da exploração dos dados.

O cenário escolhido para a manipulação da informação é a Rede Social, onde é apresentada uma estrutura dinâmica e complexa; a possibilidade de reunir enorme quantidade de informação através de um elo de comunicação com várias pessoas, e ainda carece de técnicas que melhorem os resultados e o desempenho das recomendações sociais feitas aos seus usuários tendo em vista a grande quantidade de dados a serem analisados, entre outras limitações existentes.

1.2 Motivação

O tema escolhido para a referida Dissertação de Mestrado é: “Uma Aplicação de Mineração de Dados para Recomendação Social.”

A justificativa do tema deve-se ao que foi relatado na problemática sobre o campo de Rede Social e sua complexidade na extração de informações. Tendo como motivação inicial a resolução de problemas fixados em uma extração de dados de maneira efetiva para o compartilhamento dos dados e interação entre os perfis de usuário; de forma que seja criada ao final uma aplicação que trabalhe com a recomendação social.

O tipo de Rede Social adotada para implementação da aplicação é uma Rede Social Baseada em Localização. Pois além da colaboração entre os usuários há o compartilhamento de dados geolocalizados, o que torna o resultado da recomendação social mais detalhado. Entretanto, estas Redes Sociais apresentam algumas deficiências quanto ao tratamento do conteúdo filtrado, deixando que a colaboração entre os usuários seja o único aspecto para ponderar a recomendação social.

Outro ponto importante é que a pesquisa de Mineração de Dados está alinhada ao foco do Grupo de Pesquisa de Gestão de Conhecimento do Laboratório de Sistemas Inteligentes (LSI) da Universidade Federal do Maranhão - UFMA, agregando uma contribuição aos trabalhos já existentes.

Portanto, a principal motivação deste Trabalho de Dissertação é apresentar uma aplicação que utiliza uma abordagem híbrida de Mineração de Dados para filtragem das informações em um ambiente de uma Rede Social Baseada em Localização, onde os resultados encontrados serão adotados em uma metodologia de Recomendação por *Tags*, com intuito que ao final sejam gerados resultados mais incisivos, que possam ser utilizados para a Recomendação Social e satisfaçam as necessidades usuário.

1.3 Objetivos

O objetivo deste trabalho é apresentar o desenvolvimento de técnicas de Mineração de Dados, por meio de uma aplicação para filtragem de informação em um ambiente de uma Rede Social Baseada em Localização, e gere ao final resultados que possam ser utilizados para a Recomendação Social e satisfaçam as necessidades do usuário. Além de buscar soluções para limitações que os Sistemas de Recomendação apresentam.

No sentido de alcançar este objetivo devem-se atender os seguintes objetivos específicos:

- Definir os conceitos referentes aos tópicos: Mineração de Dados, Rede Social e Sistema de Recomendação.
- Elaborar o estudo e a coleta dos dados pertinentes da Rede Social Baseada em Localização.
- Desenvolver uma metodologia de abordagem híbrida filtragem de informação com algoritmos de Mineração de Dados em Texto.
- Desenvolver uma abordagem da Recomendação por *Tags* a ser adotada ao final da filtragem de informações.
- Implementar a aplicação de Mineração de Dados para a Recomendação Social entre os usuários que contemple atributos como: localidades de interesse (restaurantes, hotéis, livrarias e etc.), conhecimento pertinente ao usuário (hobby, habilidades, comportamento) ou outras interações semelhantes.

1.4 Estrutura da Pesquisa

Este Trabalho de Dissertação está dividido em oito capítulos.

No Capítulo 2 é apresentado o referencial teórico sobre a Mineração de Dados delimitando a forma que o conhecimento é extraído por meio das técnicas e categorias de Mineração de Dados abordados.

No Capítulo 3 é esclarecido a respeito dos principais métodos ou algoritmos de Mineração de Dados utilizados atualmente, inclusive abordando alguns dos algoritmos que foram trabalhados nesta pesquisa.

No Capítulo 4 é realizado um estudo geral a respeito das Redes Sociais, ambiente onde será a aplicado a recomendação social, apresentando suas características.

Adiante, o Capítulo 5 apresenta uma análise da estrutura de um Sistema de Recomendação demonstrando suas especificidades e as limitações que ainda são encontradas atualmente.

A metodologia da pesquisa é abordada no Capítulo 6 onde é esclarecido a respeito da Recomendação Social e as Redes Sociais Baseadas em Localização apresentando pesquisas nestas áreas que ainda são recentes na computação. O Capítulo expõe também por meio da arquitetura da aplicação, a forma que a aplicação será desenvolvida e como procederá ao uso dos métodos de Mineração de Dados aplicados junto a utilização da Recomendação por *Tags* para alcançar o resultado esperado,

Por conseguinte, o Capítulo 7 demonstra uma implementação teste da aplicação proposta no capítulo anterior apresentando os resultados.

Por fim, o Capítulo 8 apresenta as considerações finais desta pesquisa e os planejamentos para o futuro.

2. MINERAÇÃO DE DADOS: UMA ABORDAGEM GERAL

Atualmente, o cenário mundial apresenta um enorme avanço tecnológico em diversas áreas (econômica, educacional, científica, institucional e etc.) e um grande volume de dados que necessita ser analisado com o auxílio de ferramentas computacionais direcionadas a cada contexto de aplicação. Para um entendimento inicial é obrigatório descrever a respeito de três aspectos da Gestão de Conhecimento: dado, informação e conhecimento.

O Dado é um “fato” distinto e objetivo que foca em um evento. Por exemplo, um tipo de dado representado nas organizações são os registros de transações, pois todas as organizações (bancos, seguradoras, Receita Federal e INSS e etc.) precisam de dados que possam representá-las ou depende dos dados para efetuar suas transações. Os dados são apenas o início para que depois sejam manipulados pela informação.

A Informação conforme destaca Peter Drucker (Davenport e Prusak, 2003) seria “dados dotados de relevância e propósito”. Simplificadamente os dados são selecionados e agrupados utilizando um critério lógico para alcançar determinado objetivo, ou seja, dados estruturados e filtrados. Atribuir esta importância a Informação requer o uso de alguns métodos:

- Contextualização: sabemos qual a finalidade dos dados.
- Categorização: conhecemos as unidades de análise ou os componentes essenciais dos dados.
- Cálculo: os dados podem ser analisados matematicamente ou estatisticamente.
- Correção: os erros são eliminados dos dados.
- Condensação: os dados podem ser resumidos para uma forma mais concisa.

2.1 Descoberta de Conhecimento em Base de Dados

O Conhecimento já recebeu mais de uma interpretação nas bibliografias, por exemplo, os autores Nonaka e Takeuchi (1997) definem conhecimento como “uma crença verdadeira justificada”.

Davenport e Prusak (2003) adotam uma definição abrangente e bem estruturada: “conhecimento é uma mistura fluida de experiência condensada, valores, informação contextual e insight experimentado, a qual proporciona uma estrutura para a avaliação e incorporação de novas experiências e informações. Ele tem origem e é aplicado na mente dos conhecedores”. A pesquisadora Angeloni (2008) declara que o conhecimento não é sinônimo de acúmulo de informações, mas um agrupamento articulado delas por meio de legitimação empírica, cognitiva e emocional, englobando a noção de “compreensão” das dimensões da realidade.

A partir destes conceitos pode-se subdividir o Conhecimento em dois campos: tácito e explícito.

O **conhecimento tácito** é aquele que não foi abstraído da prática ou em que o individuo adquiriu ao longo da vida. Este tipo de conhecimento se apresenta como mais valioso, devido a sua difícil captura, registro e divulgação, exatamente por ele estar ligado às pessoas. Ou seja, o conhecimento tácito é sutil e próprio de cada pessoa. Fica armazenado no cérebro humano aguardando o contexto adequado para tornar-se explícito. Não que dependa de uma repetição da experiência, pois poderá ressurgir num evento totalmente distinto da experiência que o originou criando uma experiência totalmente nova.

O **conhecimento explícito** (normas, manuais, legislação, códigos de conduta etc.) pode ser facilmente compartilhado na organização, não exigindo necessariamente o contato pessoal. O que existe hoje em termos de tecnologia da informação e outras automatizações de registros de uma empresa são definidos como conhecimento explícito. Estes repositórios de dados registram a experiência da organização, além de uma grande diversidade de dados sobre o seu ambiente interno (processos, históricos, profissionais e etc.) e externo (clientes, fornecedores, concorrentes e vendas). A partir dessa base, os softwares de última geração

conseguem realizar com rapidez o que o ser humano levaria muito tempo para fazê-lo: processar essa grande massa de dados e tirar dela informações relevantes.

Abaixo, segue a Figura 1 com uma arquitetura esquemática sobre o fluxo de elementos entre dados, informação e conhecimento e como é direcionado. Esta distribuição é baseada nos trabalhos do pesquisador David MacCandless: Taxonomia de Idéias[1] e contribuindo para este trabalho de dissertação.

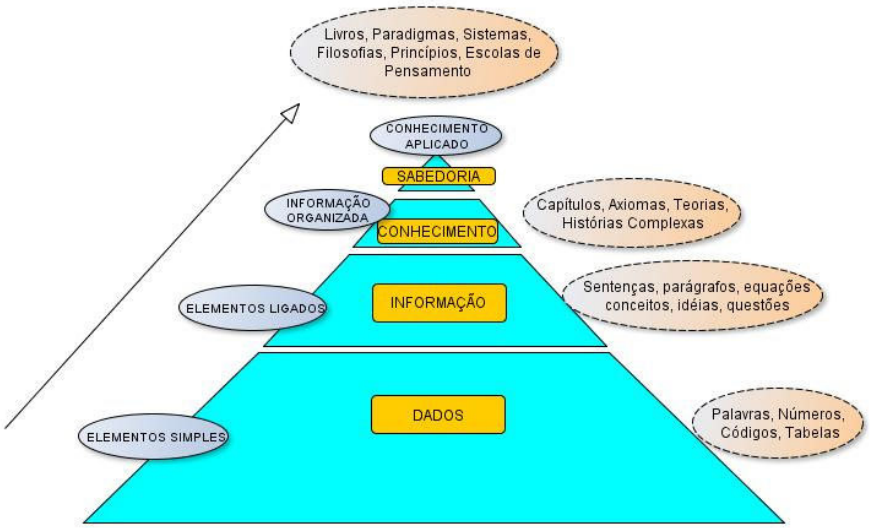

Figura 1: Arquitetura esquemática de dados, informação e conhecimento.

Para obter o estudo e manipulação do conhecimento surgiu a área de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* - KDD) que vem ganhando destaque através das comunidades científicas e ambientes organizacionais. A expressão Mineração de Dados é na verdade umas das etapas do processo de Descoberta de Conhecimento, sendo esta etapa muitas vezes confundida com o próprio processo KDD por ser a principal etapa na Descoberta de Conhecimento.

A Descoberta de Conhecimento em Bases de Dados é uma área multidisciplinar e atua em diversas áreas: Estatística, Inteligência Computacional,

[1] Trabalhos apresentados na página *Web*: http://www.informationisbeautiful.net/visualizations/ .

Reconhecimento de Padrões e Banco de Dados, Aprendizagem de Máquina, Inteligência Coletiva, e etc.

O termo KDD foi devidamente formalizado pelos autores em 1989 e teve uma das definições mais populares em 1996 por um grupo de pesquisadores (Fayyad et al., 1996a): “KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados”.

O processo de KDD é composto por várias etapas operacionais que pode ser resumida em três grupos: pré-processamento, mineração de dados e pós-processamento. Segundo o autor Goldschmidt (2005), a etapa de Pré-processamento diz respeito às funções relacionadas com captação, à organização e o tratamento dos dados, preparando os dados para o algoritmo da próxima etapa; na Mineração de Dados é realizada a busca efetiva por conhecimentos úteis no contexto da aplicação de KDD; e por fim, a etapa de Pós-processamento abrange o conhecimento que foi gerado com a Mineração de Dados.

Para um entendimento melhor do processo de KDD é necessário a compreensão dos elementos que compõem esta área. O processo KDD procura a resolução de determinado **problema**, sendo este desdobrado em três grupos: o conjunto de dados, o especialista no domínio da aplicação e os objetivos de aplicação.

O conjunto de dados é o processo de KDD que trabalha os dados de forma agrupada classificando-os em casos ou registros em uma única estrutura tabular bidimensional contendo casos e características do problema a ser analisado.

O especialista no domínio da aplicação compreende a pessoa ou o grupo de pessoas que conhece o assunto e o ambiente em que se realiza a aplicação KDD. Geralmente os especialistas detêm o conhecimento prévio sobre o problema (“*background knowledge*”).

Nos objetivos da aplicação são apresentadas as características esperadas do modelo de conhecimento a ser produzido ao final do processo; denotando as restrições e expectativas dos especialistas do domínio da aplicação acerca do modelo de conhecimento a ser gerado.

Após o problema delimitado é averiguado os **recursos disponíveis** (especialista em KDD, ferramenta de KDD e a plataforma computacional) para solucioná-lo. Convém destacar o contraste entre “ferramenta de KDD” e “plataforma computacional”. A primeira diz respeito a qualquer recurso computacional a ser utilizado para análise dos dados; sendo desde um ambiente de software até algoritmos como este fim específico de KDD. A plataforma trata dos recursos computacionais de hardware direcionados para aplicações de KDD (por exemplo, máquinas isoladas até mesmo ambientes computacionais paralelos).

Após as etapas anteriores são apresentados os resultados obtidos onde são colocados os modelos de conhecimento descobertos ao longo das aplicações de KDD. Cada modelo é avaliado de acordo com o atendimento a relação de requisitos definidos no objetivo de aplicação. Para entendimento da origem dos modelos de conhecimento é que também abrigado um “histórico” sobre como os modelos de conhecimento foram gerados e para controle de todo o processo e quando for conveniente revisar todas as ações realizadas.

Sabendo que a etapa inicial do processo de Descoberta de Conhecimento é conhecida como “Pré-processamento” o autor Goldschmidt (2005) apresenta o detalhamento das principais funções:

- **Seleção de Dados:** compreende, em essência, a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de KDD. A seleção dos dados pode ter dois enfoques distintos: a escolha de atributos ou a escolha de registros que devem ser considerados no processo de KDD.

- **Limpeza dos Dados:** abrange qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos por eles representados.

- **Codificação dos Dados:** os dados devem ser codificados para ficarem em uma forma que possam ser usados como entrada de algoritmos de Mineração de Dados. A codificação pode ser: Numérica – Categórica, que transforma valores reais em categorias ou intervalos; ou Categórica – Numérica, que representa numericamente valores de atributos categóricos.

- **Enriquecimento dos dados:** consiste em conseguir de alguma forma mais informação que possa ser agregada aos registros existentes, enriquecendo os dados, para que estes forneçam mais informações para o processo de descoberta de conhecimento. Podem ser realizadas pesquisas para a complementação dos dados, consultas a bases de dados externas, entre outras técnicas.

A etapa da Mineração de Dados propriamente dita inicia com a busca efetiva dos dados encontrados a fim de se encontrar conhecimento útil. A etapa final de Pós-processamento já contempla a análise daquilo que foi minerado na etapa anterior. Este tratamento dos dados é visto como forma de facilitar a interpretação e avaliação do homem quanto ao que foi descoberto. Abaixo segue Figura 2 com as etapas do processo de KDD.

O papel do usuário neste contexto é conduzir os processos de KDD sendo de uma participação especializada. Entretanto, para analisar a execução do processo KDD; o homem necessita utilizar sua experiência anterior, seus conhecimentos e sua intuição para interpretar e combinar subjetivamente os fatos de forma a decidir qual estratégia a ser adotada (Fayyad et. al., 1996b). O trabalho do especialista em KDD não é trivial e, portanto, sugere que o profissional não apresente apenas fundamentação teórica, mas a participação em situações reais.

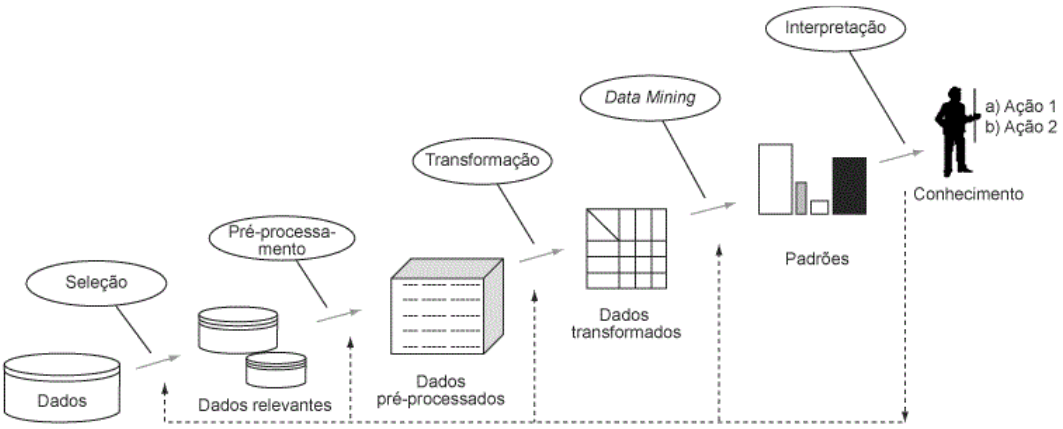

Figura 2: Etapas do processo de Descoberta de Conhecimento (Fonte: Fayyad, 1996).

2.2 Mineração de Dados

A Mineração de Dados é uma técnica conhecida por ser um processo de extração de informação sem um conhecimento prévio de uma base de dados e geralmente é utilizado para um fim especifico, de apoio a decisão ou recomendação. O pesquisador Usama Fayyad (1996) declara “é o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

Em outras literaturas encontram-se as seguintes definições:

- Mineração de Dados é o processo de proposição de várias consultas e extração de informações úteis, padrões e tendências, freqüentemente desconhecidos, a partir de grande quantidade de dados armazenada em banco de dados (Thuraisingham, 1999).

- Mineração de dados é a busca de informações valiosas em grandes bancos de dados. É um esforço de cooperação entre homens e computadores. Os homens projetam banco de dados, descrevem problemas e definem seus objetivos. Os computadores verificam dados e procuram padrões que casem com as metas estabelecidas pelos homens (Weis & Indurkhya, 1999).

- Mineração de Dados, de forma simples, é o processo de extração ou mineração de conhecimento em grandes quantidades de dados. (Han, 2001).

As técnicas de Mineração de Dados são aplicadas em várias áreas que usam do conhecimento como força motriz: empresas, indústrias e instituições de pesquisas, por exemplo. A Mineração de Dados – também chamada de *Data Mining* – define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para descrever características do passado como para predizer tendências para o futuro (Sferra, 2003).

Outra definição para Mineração de Dados seria a que propõe Braga (2005): a Mineração de Dados compreende um conjunto de técnicas para descrição e

predição a partir de grandes massas de dados. Por este motivo ela está geralmente associada a Bancos de Dados Especiais denominados *Data Warehouse.* A partir do momento em que se destacam o uso de técnicas, é abordado também a aplicação de algoritmos sobre os dados que se desejam encontrar. O que é gerado neste percurso é tido como “Modelos de Conhecimento”.

Para encontrar respostas ou extrair conhecimento útil, existem diversos métodos de Mineração de Dados disponíveis nas literaturas. Porém, para que a descoberta de conhecimento seja relevante, é importante estabelecer algumas tarefas (que apresentam nomes similares em várias outras referências): Classificação, Modelos de Relacionamento entre Variáveis, Análise de Agrupamento, Sumarização, Modelo de Dependência, Regras de Associação e Análise de Séries Temporais, conforme citação e definição feita pelo pesquisador Fayyad (1996).

A dificuldade existente no processo de Mineração de Dados é perceber e interpretar a gama de informações; distinguindo nestas informações aquilo que deve ser retido e quais ações a serem a tomadas (qual técnica a ser utilizado para alcançar sucesso na extração dos dados) em cada caso procurado. Entre os usos mais conhecidos, que podem ser observadas na Mineração de Dados são: telefonia, rede *fast-food*, educação, área médica, área financeira entre outros.

2.2.1 Técnicas de Mineração de Dados

. Entre as principais técnicas utilizadas em mineração de dados temos técnicas estatísticas, técnicas de aprendizado de máquina e técnicas baseadas em crescimento-poda-validação. Segundo Amo (2003), as técnicas podem ser classificadas em:

Análise de Regras de Associação: Uma regra de associação é um padrão da forma $X \rightarrow Y$, onde *X* e *Y* são conjuntos de valores (artigos comprados por um cliente, sintomas apresentados por um paciente, etc.). Um exemplo seria o seguinte cenário padrão de “Clientes que compram pão também compram leite”. Isto

representa uma regra de associação que apresenta um padrão de comportamento dos clientes do supermercado.

Análise de Padrões Seqüenciais: Um padrão sequencial é uma expressão da forma $< I_1,..., I_2 >$, onde cada I_i é um conjunto de itens. A ordem em que estão alinhados estes conjuntos reflete a ordem cronológica em que aconteceram os fatos representados por estes conjuntos. Assim, por exemplo, a seqüência < {carro}, {pneu, toca-fitas} > representa o padrão “Clientes que compram carro, tempos depois compram pneu e toca-fitas de carro”.

Classificação e Predição: A Classificação é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados. O modelo construído baseia-se na análise prévia de um conjunto de dados de amostragem ou dados de treinamento, contendo objetos corretamente classificados. Por exemplo, suponha que o gerente do supermercado está interessado em descobrir que tipo de característica de seus clientes serve para classificar em “bom comprador” ou “mau comprador”. Um modelo de classificação poderia incluir a seguinte regra: “Clientes da faixa econômica B, com idade entre 50 e 60 são maus compradores”. Em algumas aplicações, o usuário está mais interessado em predizer alguns valores ausentes em seus dados, em vez de descobrir classes de objetos. Isto ocorre, sobretudo quando os valores que faltam são numéricos. Neste caso, a tarefa de mineração é denominada Predição.

Análise de *Clusters* (Agrupamento): Diferentemente da classificação e predição onde os dados de treinamento estão devidamente classificados e as etiquetas das classes são conhecidas, a análise de clusters trabalha sobre dados onde as etiquetas das classes não estão denítidas. A tarefa consiste em identificar agrupamentos de objetos, agrupamentos estes que identificam uma classe. Por exemplo, pode-se aplicar análise de clusters sobre o banco de dados de um supermercado a fim de identificar grupos homogêneos de clientes, por exemplo, clientes aglutinados em determinados pontos da cidade costumam vir ao supermercado aos domingos, enquanto clientes aglutinados em outros pontos da cidade costumam fazer suas compras às segundas-feiras.

Análise de *Outliers*: Um banco de dados pode conter dados que não apresentam o comportamento geral da maioria. Estes dados são denominados *outliers* (exceções). Muitos métodos de mineração descartam estes *outliers* como sendo ruído indesejado. Entretanto, em algumas aplicações, tais como detecção de fraudes, estes eventos raros podem ser mais interessantes do que eventos que ocorrem regularmente. Por exemplo, podemos detectar o uso fraudulento de cartões de crédito ao descobrir que certos clientes efetuaram compras de valor extremamente alto, fora de seu padrão habitual de gastos.

Abaixo, segue a Figura 3 demonstrando umas das técnicas de Mineração de Dados voltadas para função de Análise de Cluster; representando três objetos que sofreram processo de clusterização através do algoritmo de *K-modes*.

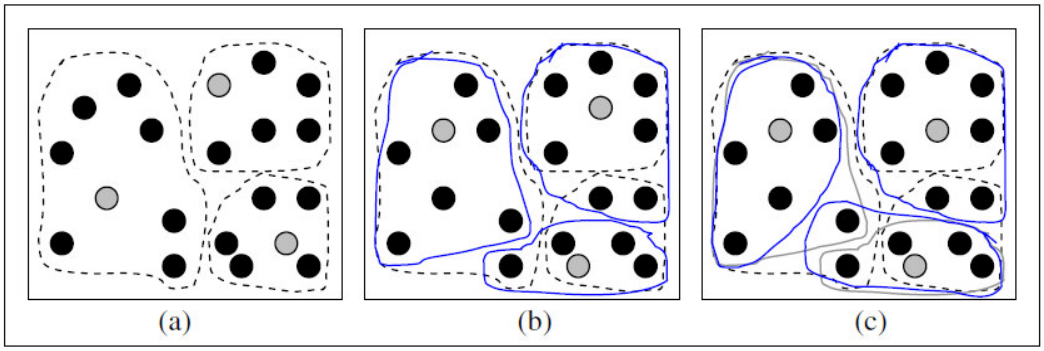

Figura 3: Representação de objetos clusterizados (Fonte: Freitas, 2008).

2.3 Mineração de Dados na *Web*

A área de Mineração de Dados na *Web* visa a descoberta de conhecimento em bases de dados localizados em um ambiente complexo com uma grande quantidade de documentos heterogêneos e um dinamismo crescente no fluxo dos dados que tramitam na *Web*. Apresenta três enfoques para sua aplicação: Mineração de Conteúdo, Mineração de Estruturas da *Web* e Mineração de Uso da *Web*.

A Mineração na *Web* pode ser definida, de forma simples, como a utilização de técnicas de Mineração de Dados para a recuperação automática, extração e avaliação de informação para a descoberta de conhecimento em documentos e

serviços da *Web*. Abaixo segue a Figura 4 para um melhor entendimento da forma que os dados são minerados pela *Web* apresentando uma estrutura de Mineração na *Web* e a distribuição de suas tarefas.

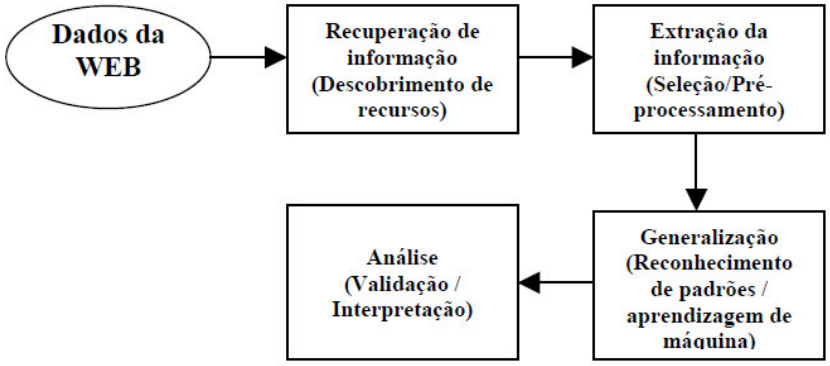

Figura 4: Etapas das Sub-Tarefas na Mineração na *Web* (Fonte: Pal, 2000).

A categoria de **Mineração de Conteúdo na *Web*** é o processo de extração de conhecimento em documentos e seus metadados (informações, autores, descrição, palavras-chave, etc.). Por exemplo: documentos textuais (páginas de texto, HTML, listas de discussão, grupos de usuários, *blogs,* etc.) e também a mineração de dados multimídia. A categoria de **Mineração de Estruturas da *Web*** é um processo de KDD voltado para a organização da *Web*, sobretudo, a organização da ligação entre os documentos na *Web*.

Por fim, a categoria de **Mineração de Uso da *Web*** trabalha com a análise dos dados coletados através de um acesso a documentos na *Web* (em particular os *logs*), com a finalidade de descobrir padrões de acesso a sites ou um conjunto de informações que possa modelar o comportamento do usuário e prover melhoras na experiência do usuário.

Ao longo do tempo, já foram realizadas algumas pesquisas de grande relevância e importância para a fundamentação teórica no campo de Mineração de Dados na *Web,* a partir destes trabalhos foram apresentadas evoluções e problemas que ainda carecem de soluções mais simples e eficazes. A Tabela 1 apresenta um quadro-resumo de alguns trabalhos divididos por temas ou tópicos de pesquisa voltados às categorias citadas anteriormente. As três categorias não são isoladas em si, pois seus trabalhos apresentam padrões de associação em determinadas

situações. Por exemplo, em alguns estudos foram utilizados dados de conteúdo dos documentos e das ligações entre documentos para tarefas específicas de Mineração de Dados; em outros momentos usam *logs* de servidores juntamente com as estruturas correspondentes dos sites para um melhor efeito na caracterização dos padrões de acesso dos usuários e até na identificação de perfis destes usuários. A natureza destes dados irá ditar qual idéia a ser sugerida para aplicar determinada categoria na Mineração de Dados para *Web*.

Tabela 1: Trabalhos nas três categorias de Mineração de Dados na *Web*.

Mineração de Conteúdo na *Web*	**Mineração de Estruturas da *Web***	**Mineração de Uso da *Web***
Análise de blogs através do uso de métricas para caracterização de tópicos (Hayes, 2007).	Apresentam métodos para monitoramento de tópicos na *Web* e para o estudo de co-visibilidade de tópicos, usando contagem de hits de sites de busca e redes semânticas, mostrando exemplos reais de aplicação (Kiefer, 2006).	Um sistema automatizado que categoriza usuários de um servidor na *Web* através de análise de agrupamentos, e que apresenta desempenho e precisão melhor do que outros sistemas existentes (Chi, 2003).
Criação de algoritmos para uso em um sistema de recomendação para blogs com conteúdo similar (Abbassi e Mirrokni, 2009).	Incorporando informações sobre o conteúdo de dois documentos na *Web* conectados por hiperlinks (Utard e Fürnkranz, 2006)	Técnicas de mineração de logs para caracterização de padrões de navegação para extrair as sessões de navegação de usuários para processamento (Berendt, 2003).
Melhoras dos resultados de busca por documentos na *Web* através do agrupamento dos documentos por frases que contém as palavras procuradas (Yang e Rahi, 2003).	Utilização de informações de relações entre blogs e classificá-los através de uma abordagem de rotulação de grafos de forma semi-supervisionada (Bhagat, 2009).	Apresentam um modelo de mineração de sites que usa dados textuais entrados nos sistemas de busca no próprio site para identificar vários tipos de resultados (Baeza-Yates e Poblete, 2006).
nálise de algoritmos de recomendação baseados em descoberta de correlações (Geyer-Schulz e Michael Hahsler, 2003).	Algoritmo de mineração de estruturas na *Web* é o *PageRank*. Avalia páginas baseado na quantidade de ligações a ela feitas por outras páginas consideradas importantes (Brin, 1998).	Técnica para personalizar resultados de um sistema de buscas na Internet usando interesse pessoal dos usuários, representado através de seus marcadores (bookmarks) que indicam interesses em páginas e tópicos (Kim e Cham, 2006).

2.3.1 Mineração de Conteúdo na *Web*

Uma definição simples para este tipo de Mineração de Dados é que esta categoria se trata da Descoberta de Informações a partir do conteúdo web: dados, documentos, serviços e similares. O conteúdo na *web* que se fala, vai muito mais além que palavras, ou ainda documento textual, pode ser apresentado na forma de: áudio, vídeos, metadados atrelados a simbologia específica e dados como *hiperlink.* Ou seja, os dados de texto da *Web* podem ser classificados em três tipos: dados não estruturados (texto livre), sem dados estruturados (HTML) e dados totalmente estruturados (tabelas e bases de dados).

O pesquisador Colley (1997) definiu uma categorização para a Mineração de Conteúdo na *Web* detalhada na Figura 5. Estas categorias se dividem em duas: Abordagem Baseada em Agentes e Baseada em Banco de Dados. O autor Pal (2000) descreve que a primeira categoria envolve o desenvolvimento de sofisticados sistemas de Inteligência Artificial que podem atuar de forma autônoma ou semi-autônoma em nome de um usuário em particular, para descobrir e organizar as informações baseadas na *Web*. Esta categoria por sua vez, se subdivide em três tipos: filtragem de informação, agentes de busca inteligente e agentes de web personalizados. A abordagem de Banco de Dados se concentra em técnicas para organizar dados semi-estruturados na web para coleta de recursos mais estruturados e utiliza mecanismos de consulta de Banco de Dados junto as técnicas de Mineração de Dados para analisá-los. Esta categoria se subdivide em dois campos: Bases de Dados de vários níveis e Sistemas *Web* de consulta.

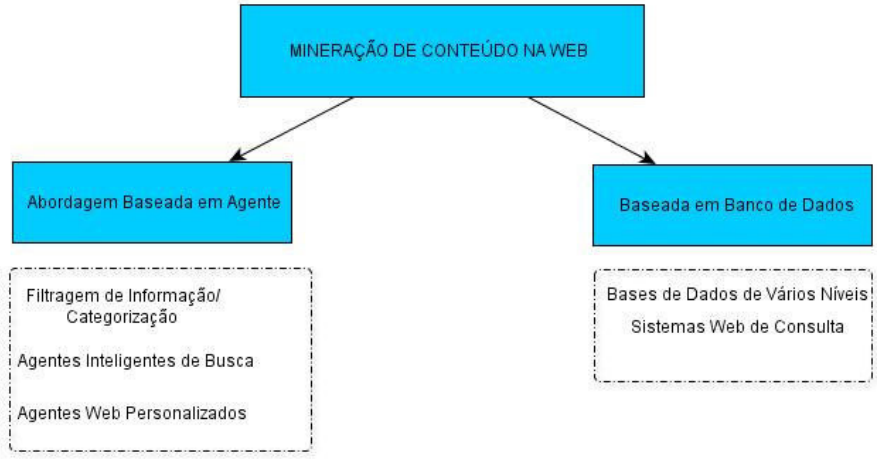

Figura 5: Taxonomia da Mineração de Conteúdo (Fonte: Colley, 1997)

O campo da Mineração de Conteúdo Multimídia apresenta um estudo a parte, umas principais referências significativas foram conduzidas por Kosh (2003) e Zaiane (1998). O primeiro apresenta uma literatura sobre este campo e suas aplicações quando a armazenagem de dados multimídia. O segundo apresenta um estudo sobre algoritmos que trabalham com dados multimídia gerando uma aplicação.

Analisando o tratamento dos dados não-estruturados há um campo de pesquisa que vem ganhando destaque nos últimos anos (sendo um dos focos deste trabalho de dissertação): a Mineração de Texto ou Mineração de dados em Texto (ou também chamada de *Knowledge Discovery in Texts*). Uma das causas que possibilitam o crescimento da pesquisa na Mineração de Dados em Texto é a possibilidade de extrair os dados não-estruturados por meio de marcações semânticas, *hiperlinks* ou algum outro vínculo que permita agrupar as informações textuais.

2.3.2 Mineração de Estruturas da *Web*

Analisar a Mineração de Estruturas da *Web* remonta a compreensão de como são formuladas as estruturas que organizam as informações textuais que a Mineração de Texto se preocupa em responder. O que permite interligar estas informações textuais são os vínculos de hipertexto.

A princípio deve-se compreender o que seria a representação da *Web.* Esta representação assemelha-se com a teoria dos grafos, em que os “nós” representam páginas, e as setas entre os pares de nós representam vínculos entre as páginas. Este tipo de representação da *Web* apresenta uma forte semelhança com a teoria modernas das redes sociais desenvolvida nos trabalhos de Stanley Milgram (Kumar, 2002).

O experimento de Milgram descrevia o seguinte roteiro: alguns associados em Omaha, Nebraska, enviavam uma carta a outro associado de Boston. As cartas só poderiam ser enviadas para aqueles que soubessem o primeiro nome e os mesmos somente poderiam reenviar a carta para aqueles que eles também sabiam o primeiro nome. A idéia era fazer a carta chegar ao associado com um menor número de “saltos”. O número médio de saltos era de seis e segundo Kumar (2002), criou-se um folclore nos Estados Unidos a respeito do número médio de cartas entregues com sucesso ao longo do caminho: que duas pessoas que trocavam cartas estavam interligadas em uma rede social de “6 graus de separação”.

As pesquisas neste campo da mineração na *web*, através das semelhanças entre o domínio *web* e as redes sociais, propuseram técnicas de Mineração de Dados voltadas para a manipulação do conhecimento. Entre algumas técnicas que se pode citar como de suma importância e pioneiras são: o HITS algoritmo (Kleinberg,1998) e o algoritmo *PageRank* (Brin, 1998).

O primeiro algoritmo citado de HITS criado por Kleinberg aborda a pesquisa do tema *hyperlinked* induzido (HITS) através da identificação de dois tipos de paginas *Web:*

- “autoridades-páginas” que representam a autoridade em fontes de informação para a consulta, e
- “centros de recursos”, listas contendo indicações sobre cada tópico.

É feita uma coleta de dados das páginas de um conjunto raiz de páginas de um motor de busca baseado em texto. O algoritmo HITS expande a raiz definida em um conjunto base, acrescentando todas as páginas que estão ligadas a ele para as páginas em o conjunto raiz. A idéia é a de assegurar que o conjunto de base conterá as melhores páginas para a consulta, mesmo que o conjunto raiz não apresente. O algoritmo *PageRank* trabalha com uma abordagem de análises de links para obter um *pagerank*, um ranking de consulta independente de todas as páginas.

A principal vantagem deste algoritmo vem da forma que trabalha sua ordenação estática, esta permite que as páginas que contenha um dado de termo de consulta possam ser recuperadas usando um método tradicional com base em: um texto e um indexador para ser exibido em uma ordem *pagerank*. O *PageRank* tornou-se o algoritmo base do motor de busca do *Google*, porém não deve ser confundido com todo o processo de busca, mas apenas um dos componentes.

2.3.3 Mineração de Uso da *Web*

Segundo o autor Kosala (2000) a Mineração de Uso da *Web* preocupa-se em prever o comportamento do usuário através de técnicas voltadas para interação do usuário e da *web*. O autor Cooley (1999) detalha a definição deste tipo de mineração de dados como: uma aplicação de técnicas de mineração de dados para Web com grandes repositórios de dados preocupando-se com o layout e navegação pela *web*. Alguns dos algoritmos de Mineração de Dados neste campo que são normalmente utilizados: associação por geração de regra, a geração de padrão seqüencial e *clustering*.

Geralmente os dados trabalhados nesta categoria de mineração da *web* são direcionados para: *logs* de servidores web e servidores Proxy; *logs* de navegadores, perfis de usuário, arquivos temporários, pasta de favoritos, consultas do usuário,

cliques de mouse e qualquer outro dado gerado pela interação do usuário com a *web*. As aplicações da Mineração de Uso da *Web* podem ser classificadas em duas categorias principais: aprendizado de perfil de usuário e aprendizado de padrões de navegação do usuário. Os sites de comércio eletrônicos e algumas redes sociais utilizam em grande parte este tipo de categoria de mineração na *web.*

A página *web* de comércio eletrônico é um exemplo bem clássico quanto a previsão do tipo de comportamento de usuário e serve como uma forma de apresentar recomendações e soluções ao que este usuário busca pela *web*. Este tipo de recomendação surge devido ao emprego da similaridade entre perfis, ou ainda perfis e itens, para filtrar os resultados mais expressivos ao que o usuário busca inclusive indiretamente. Abaixo segue a Figura 6 com um exemplo de arquitetura da Mineração de Usuário *Web* onde o diferencial é a pesquisa direcionada a "sessão" do usuário pela *Web.*

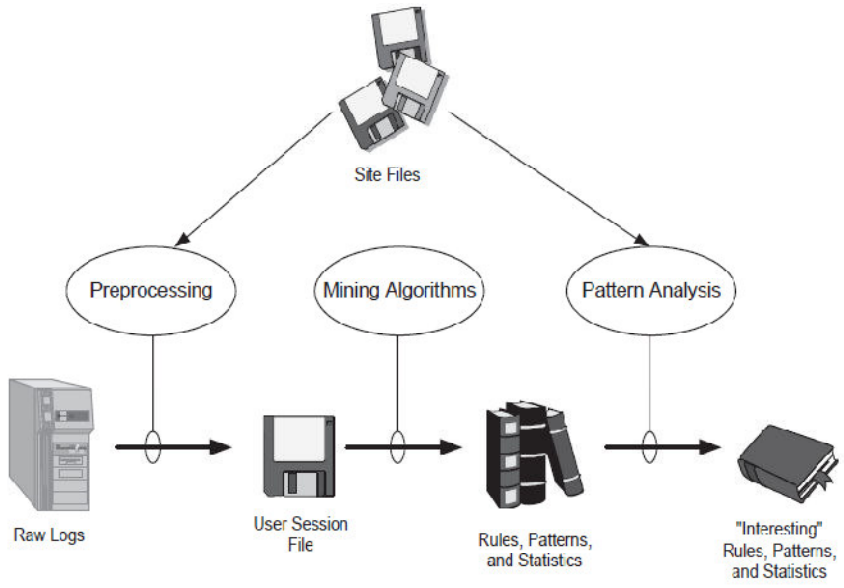

Figura 6: Arquitetura da Mineração de Uso na *Web* (Fonte: Srisvastava, 2000).

2.4 Mineração de Dados em Texto

A Mineração de Dados em Texto faz o tratamento dos dados desestruturados da Web sendo chamado pelo termo KDT (*Knowledge Discovery in Texts*) que pode

ser definido como o processo de extrair padrões interessantes e não-triviais, a partir de documentos textuais (Tan A.-H, 1999).

Outra definição para Mineração de Dados em Texto é: o estudo e a prática de extrair informação de textos usando os princípios da lingüística computacional (Sullivan, D., 2000). O pesquisador Hearst (1999) apresenta esta mineração como uma análise de dados exploratória; sendo um método para apoiar pesquisadores a derivar novas e relevantes informações de uma grande coleção de textos. É um processo parcialmente automatizado onde o pesquisador ainda está envolvido, interagindo com o sistema.

O que a diferencia Mineração de Textos da Mineração de Dados é que a primeira obtém conhecimento através de bases textuais (documentos em linguagem natural) enquanto a Mineração de Dado em si, preocupa-se com as bases estruturadas como sistemas gerenciadores de Banco de Dados. Convém destacar que em alguns casos a própria mineração feita em textos pode resultar na transformação do texto em dados estruturados. Os dois processos de mineração não são exclusivamente distintos ambos podem trabalhar com as mesmas técnicas de extração de dados como: Aprendizagem de Máquina.

Atualmente, os investimentos e estudos são direcionados para a Mineração de Textos voltados para o ambiente *Web*. Neste seguimento surgiu uma nova área de pesquisa voltada para a Mineração de Dados textuais: Mineração de Texto da Web (*Web Text Mining*). A pesquisadora Bastos (2006) destaca que este campo de estudo está dividido em três categorias:

- Mineração de Conteúdo da *Web* (*Web Content Mining*) – descreve a descoberta de informação útil em conteúdos de páginas ou documentos na Web. É possível encontrar bibliotecas digitais acessíveis na Web, como também empresas que expõem informações a respeito de oportunidades de negócios ou serviços eletronicamente, disponibilizando esses dados para seus clientes, funcionários e parceiros, através de interfaces de navegação (*browsers*). Algumas informações de conteúdos de páginas e sites estão escondidas, não permitindo o acesso para usuários não autorizados, e, portanto, não podendo ser exploradas.

- Mineração de Estrutura da *Web* (*Web Structure Mining*) – extrai o modelo que descreve as conexões entre páginas de um site. Este modelo é baseado na topologia de *hyperlinks*, com ou sem a descrição dos *links*. A partir deste modelo é possível categorizar páginas Web, bem como gerar informações a respeito das semelhanças e relacionamentos existentes entre sites diferentes.

- Mineração de Usuário da *Web* (*Web Usage Mining*) – esta categoria preocupa-se com as técnicas de predição de comportamento de usuários, enquanto o usuário interage com a Web. A informação que é pesquisada nesta área de interesse é chamada de dado secundário, e é obtido a partir da navegação do usuário no site. Estes dados incluem *logs* de acesso armazenados no servidor Web, *logs* do servidor de *proxy*, *logs* do navegador *Web*, perfis do usuário, resultados de consultas efetuadas pelo usuário, *cookies*, cliques do mouse, arquivos temporários e outros similares.

O principal elemento de uma Mineração de Dados Textuais é a coleção de documentos, sendo que os cenários da *Web* onde se encontram estes documentos podem ser estáticos ou dinâmicos. Em ambiente estático os documentos permanecem inalterados tanto em seus conteúdos quanto nos elementos que a compõem. No entanto, o que se encontra hoje é uma enorme quantidade conteúdos dinâmicos, ou seja, elementos que podem ser alterados ou excluídos; resultado em um cenário complexo para Mineração de Dados em Texto neste campo.

O cenário dinâmico da *Web* também pode ser caracterizado como um ambiente da Rede Social, onde se demonstra a dificuldade da Mineração de Texto neste campo. Os documentos da Rede Social em sua maioria fornecem informações em grande escala em sua estrutura, do que a existente no próprio texto. Existem documentos com enormes graus de formatação (*tags*) estrutural ou visual; onde metadados podem ser inferidos (documentos semi-estruturados); e depois informações podem ser extraídas pela forma estrutural deste tipo de documento; por exemplo, documentação textual presente na linguagem XML (Feldman & Sanger, 2007).

Para alcançar a compreensão correta destes dados textuais deve-se percorrer por algumas áreas que trabalham com a análise da linguagem natural como: o agrupamento de palavras, detecção de entidades, a segmentação de

sentenças e uso de métricas de similaridade. Portanto, o KDT trabalha com técnicas para extração do conhecimento das bases textuais, como por exemplo: algoritmo TF-IDF, similaridade de cosseno, modelos de espaço vetorial, métricas de similaridade de agrupamento, técnicas de *stemming*, métodos probabilísticos e etc.

Assim como a Mineração de Dados convencional, as técnicas de Mineração de Dados em Texto estão representadas em uma estrutura de Descoberta de Conhecimento dos Dados. A Figura 7 ilustra as etapas do KDD com a mineração de dados textual agindo como realizador da extração de conhecimento; além de apresentar em cada etapa outras metodologias de pesquisa que contribuem para a descoberta de conhecimento (*Crawling,* Linguagem Natural, Recuperação de Informação e Interpretação dos Dados).

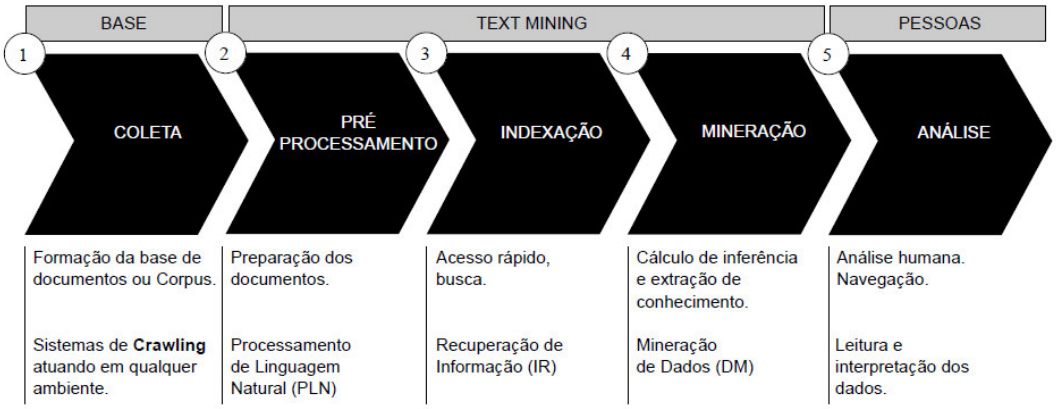

Figura 7: Etapas da Descoberta de Conhecimento de dados textuais (Fonte: Aranha, 2007).

O próximo Capítulo abordará os Métodos utilizados para realizar as operações de Mineração de Dados, sobretudo os algoritmos que foram utilizados para o desenvolvimento da metodologia a ser aplicado na Filtragem de Informação e manipulação dos dados textuais desta pesquisa da Dissertação.

3. MÉTODOS DE MINERAÇÃO DE DADOS

No capítulo anterior é delimitado um tópico onde são descritos as principais técnicas de Mineração de Dados: Análise de Regras de Associação, Análise de Padrões Seqüenciais, Classificação e Predição, Análise de *Clusters* e Análise de *Outliers*. Na utilização destas técnicas são utilizados métodos ou algoritmos de Mineração de Dados (ou Métodos de Mineração de Dados) que trabalham junto a Filtragem da Informação; seja esta uma Filtragem Baseada em Conteúdo, uma Filtragem Baseada na Colaboração ou ainda uma Filtragem Híbrida – que mescle as duas primeiras abordagens. A natureza dos dados junto a forma que ele é manipulado e distribuído no ambiente *web* (em um ambiente estático ou dinâmico) definirá qual a melhor estratégia a ser adotada para a filtragem das informações e qual algoritmo a ser escolhido.

Neste capítulo são apresentados modelos onde atuam as técnicas de Mineração de Dados Textual e os principais algoritmos adotados na linha de pesquisa que se encontra dentro da problemática que o trabalho de Dissertação propõe a solucionar.

3.1 Modelos para Mineração de Dados em Texto

Os modelos clássicos de Recuperação da Informação ou da Mineração de Dados em Texto apresentam estratégias para extração de informação relevante em documentos, sobretudo do ambiente *web,* e dividem-se em: modelo booleano, modelo vetorial e modelo probabilístico (Baeza-Yates, 1998).

3.1.1 Modelo Booleano

O modelo booleano baseia-se em álgebra booleana da teoria dos conjuntos. Dada uma consulta formada por uma expressão booleana sobre as ocorrências de determinados termos, operações como interseção e união são realizadas nos conjuntos de documentos do sistema, sendo o conjunto resultante tomado como resultado.

Este modelo tem como principal vantagem sua simplicidade e sua facilidade de implementação. Entretanto, sua rigidez é um problema, pois a inexistência de um dos termos de uma consulta já descarta um documento, provocando respostas hora com um número grande, hora com um número pequeno de documentos.

3.1.2 Modelo Vetorial

Os modelos vetoriais podem ser vistos como uma extensão do modelo booleano. Para permitir que os resultados das consultas incluam documentos que não similares exatamente ao padrão da consulta, pesos não binários são atribuídos a cada termo. Neste modelo, a resposta de uma consulta realizada lista os documentos de acordo com o grau de similaridade em relação à consulta.

O Modelo de Espaço Vetorial, também chamado de Modelo Vetorial, criado por Salton para ser utilizado em um Sistema de Recuperação de Informação chamado SMART, representa documentos e consultas como vetores de termos (Belkin, 1992).

Os termos de consulta e documento são atribuídos pesos que significam o tamanho e a direção de seu vetor de representação. Entre as técnicas mais importantes para calcular o peso de cada termo em um documento, o TF-IDF (*Term Frequency Inverse Document Frequency*) se apresenta como o mais difundido (Salton, 1988). Onde *TF* é freqüência de um termo em relação ao documento

especifico e *IDF* é a freqüência inversa de um termo quanto ao corpus do documento. Abaixo ambas as equações são apresentadas:

$$tf_{t,d} = freq_{t,d} \tag{1}$$

$$idf_t = \frac{N}{n_t} \tag{2}$$

Onde,

- frequência de um termo *tf* é o número de vezes que um determinado termo *t* aparece no texto de um documento *d*.
- *N* é o número de documentos em um corpus e n_t como o número de documentos com o termo *t.*

Nesta técnica os documentos são vistos como vetores de palavras inseridos em um espaço vetorial. A multiplicação dessas duas freqüências produz uma pontuação que customiza ambas as freqüências e procura sanar as deficiências de cada uma, por exemplo: uma palavra-chave que aparece muitas vezes em um corpus não pode caracterizar o documento como relevante ou não. Abaixo segue a descrição do algoritmo TF-IDF:

$$w_{t,d} = tf_{t,d} \times idf_t \tag{3}$$

Entre as vantagens do Modelo Vetorial há o bom desempenho em achar documentos semelhantes à descrição da consulta e permitir a criação de um *ranking* dos documentos de acordo com esta similaridade. O modelo apresenta um bom comportamento em documentos variados.

Uma limitação desta técnica seria que um documento relevante poderia não ter um termo de consulta.

3.1.3 Modelo Probabilístico

Este modelo é baseado no princípio de ordenação probabilístico (*Probability Ranking Principle*). Neste modelo, busca-se saber a probabilidade de um documento *D* ser ou não ser relevante para uma determinada consulta *Q* (Santos, 2008). O Teorema de Bayes é a principal ferramenta do Modelo Probabilístico.

O modelo probabilístico tenta estimar a probabilidade de um usuário achar um documento relevante para uma determinada consulta realizada, sendo apenas a consulta e a representação do documento levados em consideração no cálculo das probabilidades (Robertson, 1976).

Dada uma consulta, o conjunto de documentos retornados deve maximizar a probabilidade de relevância para o usuário. Como vantagem, este modelo permite consultas com documentos ordenados pela probabilidade de relevância garantindo um bom desempenho. Como desvantagem, este método ignora as freqüências dos termos em um documento e assume a independência entre eles. E depende da precisão das estimativas de probabilidade.

3.2 Métodos Baseados em Redes Neurais

Os modelos de Redes Neurais podem ser implementados nas cinco Técnicas de Mineração de Dados relatados anteriormente, onde a sua topologia segue o seguinte padrão: a camada de entrada do modelo neural recebe os dados do pré-processamento, a Rede Neural processa estes dados e gera uma saída que dependerá da função da aplicação; ou seja, as Redes Neurais com Aprendizado Supervisionado a saída do modelo vai corresponder ao atributo objetivo do problema.

Em termos mais práticos as Redes Neurais são ferramentas estatísticos não-lineares de modelagem de dados. Eles podem ser utilizados para modelar as relações complexas entre entradas e saídas, ou para encontrar padrões em dados.

As Redes Neurais são constituídas por três peças: a arquitetura ou modelo, o algoritmo de aprendizagem; e as funções de ativação. São programadas ou "treinadas" para armazenar, reconhecer e associativamente recuperar padrões ou de banco de dados entradas, para a solução de problemas de otimização combinatória; ao filtrar o ruído a partir de dados de medição, para controlar problemas mal definidos, em resumo, para estimar funções amostradas quando não sabemos a forma das funções. As duas habilidades reconhecimento de padrões e função estimação tornam as Redes Neurais propícia para a Mineração de Dados (Singh, 2005).

Segundo Goldschmidt (2005), o algoritmo de aprendizado pode estimar o erro (ou distância) entre a saída produzida pela rede e a saída desejada e através do erro calculado o algoritmo ajusta os pesos das conexões da rede a fim de tornar a saída real tão próxima da saída desejada. Por isso, o uso das Redes Neurais nos trabalhos voltados ao reconhecimento de padrões e em particular para as técnicas de Classificação e Predição da Mineração de Dados. O algoritmo de aprendizagem não supervisionado é adequado para as técnicas de Mineração de Dados que trabalhem com o agrupamento de dados como a Análise de *Clusters*.

O algoritmo *Back-Propagation* ou Algoritmo de Retropropagação de Erro é conhecido como um algoritmo supervisionado, onde diminui a função de erro entre a saída gerada pela rede neural e a saída real desejada através do método gradiente descendente. Os neurônios são organizados em camadas e enviam seus sinais de frente e em seguidas seus sinais de erro são propagados para trás. A propagação de volta do algoritmo utiliza aprendizagem supervisionada, o que significa que fornecem exemplos de entrada e saída do algoritmo para que rede calcule; em seguida, o erro (diferença entre os resultados reais e esperados) é calculado até que seja reduzido por meio deste treinamento da rede neural (Singh, 2009).

Os Mapas Auto-Organizáveis são conhecidos por ser um algoritmo não supervisionado, onde seu treinamento é baseado em uma forma de competição entre os elementos processadores (*Competitive Learning*). Segundo Goldschmidt (2005), consiste em uma forma de aprendizado que divide o conjunto de padrões de entrada em grupos inerentes aos dados, considerando em sua abordagem mais

simples, que os neurônios de saída competem entre si, resultando em apenas um neurônio vencedor (com maior ativação).

3.3 Método K- Vizinhos mais Próximos (*K-Nearest Neighbors*)

Este método é muito utilizado nas técnicas de Mineração de Dados que envolvem a Classificação e Predição, e foi proposto inicialmente por Cover (1967). O algoritmo calcula a distância, através de uma métrica, dos vizinhos mais próximos de um novo usuário x em uma base de referência k ou de um usuário com um documento (Santos, 2008).

De todos os algoritmos, o Método do Vizinho mais Próximo (ou *K-Nearest Neighbors,* ou ainda K-NN) é amplamente utilizado como classificador de texto por causa da sua simplicidade e eficiência. Sua fase de treinamento consiste em armazenar todos os exemplos de treinamento como classificador, adiando a decisão sobre como generalizar além dos dados de treinamento, até que cada nova instância de consulta seja encontrada (Tan, 2006 apud Sebastiani, 2002).

O algoritmo se apresenta no cenário que caracteriza a utilização deste método com: uma base de dados de um problema envolvendo a Técnica de Mineração de Dados de Classificação e cada novo registro a ser classificado. O autor Goldschmidt (2005) destaca os seguintes passos a serem executados:

a) Cálculo da distância do novo registro a cada um dos registros existentes na base de dados utilizando uma métrica de distância (Distância Euclideana, Distância de Hamming e Distância de Minkowski);

b) Identificação dos k registros da base de referência que apresentam menor distância em relação ao novo registro (mais similares);

c) Apuração da classe mais freqüente entre os k registros identificados no passo anterior;

d) Fazer uma comparação da classe apurada com a classe real computando o erro ou acerto do algoritmo, só deve ser efetuado este processo quando as

classes dos novos registros são conhecidas e deseja-se avaliar o desempenho do método *K-NN* na base de dados.

Portanto, sinteticamente o Método K-NN apresenta duas fases: a primeira é a determinação dos vizinhos mais próximos, e a segunda é a determinação das classes utilizando os vizinhos. Abaixo, a Figura 8 apresenta a utilização do algoritmo dos Vizinhos mais Próximos, em um contexto simples, onde pode se identificar um padrão desconhecido x_d entre padrões da classe "+" e padrões da classe "○". Usando o algoritmo K-NN, descrito nessa seção, o padrão x_d é classificado como sendo da classe "+", que é a classe do seu vizinho mais próximo conforme delimitado pela área tracejada.

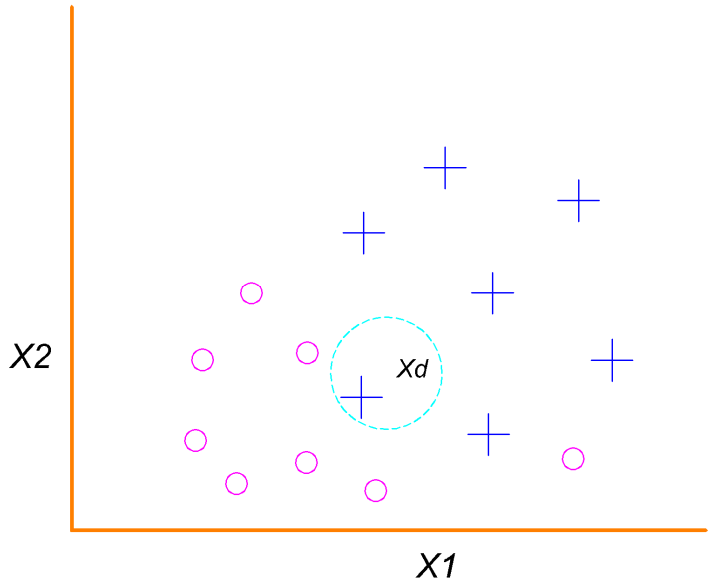

Figura 8: Representação Gráfica do algoritmo *K-NN*.

O diferencial deste Método é a possibilidade de filtrar itens baseados na predominância de k vizinhos mais próximos. Onde k é igual ao número de vizinhos mais próximos e os vizinhos mais próximos são os usuários que possuem maior valor de similaridade, efetuando uma Generalização e Classificação (Santos, 2008).

3.4 Classificador Bayesiano Ingênuo

O algoritmo também chamado de Teorema de Bayes está relacionado aos cálculos que envolvem probabilidade condicional (Goldschmidt, 2005). É aplicável em Técnicas de Classificação e Predição.

Sejam $X(A_1, A_{2,...,} A_n, C)$ um conjunto de dados, $C_1, C_{2,...,} C_k$, as classes do problema (valores possíveis do atributo *C*) e *R* um novo registro que deve ser classificado. Sejam ainda $a_1, a_{2,...,} a_n$ os valores que *R* assume em *X*. O Classificador Bayesiano Ingênuo possui dois passos:

a) Calcular a probabilidade $P(C = C_i / R), i = 1,2,...,k.$

b) Indicar como saída do algoritmo a classe C_j tal que $P(C = C_j / R)$ seja máxima. O problema reduz-se, portanto, ao cálculo das probabilidades condicionais $P(C = C_i / R), i = 1,2,...,k.$ Sabe-se que $P(C = C_i / R)$ pode ser reescrito como:

$$P(C = C_i / A_1 = a_1 e\ A_2 = a_2 e \ldots \text{e}\ A_n = a_n)$$

Por outro lado, pelo Teorema de Bayes, como *P(A/B) = (P(B/A)*P(A))/P(B),*

$$P(C = C_i / A_1 = a_1 e\ A_2 = a_2 e \ldots \text{e}\ A_n = a_n) =$$
$$(P(C = C_i / A_1 = a_1 e\ A_2 = a_2 e \ldots \text{e}\ A_n = a_n) * P(C = C_i)) /$$
$$P(A_1 = a_1 e\ A_2 = a_2 e \ldots \text{e}\ A_n = a_n)$$

O denominador na igualdade mostrado anteriormente será sempre o mesmo, independente da classe para qual a probabilidade esteja sendo calculada. Assim sendo, para fins de comparação entre as probabilidades, o denominador $P(A_1 = a_1 e\ A_2 = a_2 e \ldots \text{e}\ A_n = a_n)$ pode ser desprezado do cálculo. Desta forma, a expressão reduz-se para:

$$P(C = C_i / A_1 = a_1 e\ A_2 = a_2 e \ldots \text{e}\ A_n = a_n) = P(C = C_i /$$
$$A_1 = a_1 e\ A_2 = a_2 e \ldots \text{e}\ A_n = a_n) * P(C = C_i)$$

O nome ingênuo no título do método decorre da premissa assumida pelo algoritmo de que atributos serão sempre independentes entre si, o que tem, em muitos casos, não deverá ocorrer. Da teoria das probabilidades, se dois eventos A e B são independentes, então *P(AB) = P(A)*P(B).* Assim sendo a igualdade acima pode ser descrita da seguinte forma:

$$P(C = C_i / A_1 = a_1 e\ A_2 = a_2 e \ldots \text{e}\ A_n = a_n) = P(A_1 = a_1 / C = C_i) *$$
$$P(A_2 = a_2 / \text{C} = C_i) * \ldots * P(A_n = a_n / \text{C} = C_i) * P(\text{C} = C_i)$$

Para ilustrar esta aplicação do teorema, o exemplo de um atributo “Jogar Tênis” que tem como objetivo a classificação e apresenta dois problemas: “Jogar Tênis = Sim” e “Jogar Tênis = Não”; abaixo na Tabela 2 é descrito um banco de dados do exemplo de aplicação do Teorema de Bayes.

Os Atributos A_i são Aparência, Temperatura, Umidade e Vento. Se houver o questionamento: “Devo ou não jogar tênis em dia ensolarado, quente, de alta umidade e vento fraco?”

Aplicado o Classificador Bayesiano Ingênuo, temos:

P(Jogar=Sim | ensolarado, quente, alta umidade, vento fraco) = P(ensolarado | Jogar=Sim) * P(quente | Jogar=Sim) * P(alta umidade | Jogar=Sim) * P(vento fraco | Jogar=Sim) = **0,0071**

P(Jogar=Não | ensolarado, quente, alta umidade, vento fraco) = P(ensolarado | Jogar=Não) * P(quente | Jogar=Não) * P(alta umidade | Jogar=Não) * P(vento fraco | Jogar=Não) = **0,0274**

A resposta para a indagação acima seria “Jogar=Não”.

Tabela 2: Base de Dados para o atributo "Jogar Tênis".

Aparência	Temperatura	Umidade	Vento	Jogar Tênis
Ensolarado	Quente	Alta	Fraco	Não
Ensolarado	Quente	Alta	Forte	Não
Nublado	Quente	Alta	Fraco	Sim
Chuvoso	Moderado	Alta	Fraco	Sim
Chuvoso	Fresco	Normal	Fraco	Sim
Chuvoso	Fresco	Normal	Forte	Não
Nublado	Fresco	Normal	Forte	Sim
Ensolarado	Moderado	Alta	Fraco	Não
Ensolarado	Fresco	Normal	Fraco	Sim
Chuvoso	Moderado	Normal	Fraco	Sim
Ensolarado	Moderado	Normal	Forte	Sim
Nublado	Moderado	Alta	Forte	Sim
Nublado	Quente	Normal	Fraco	Sim
Chuvoso	Moderado	Alta	Forte	Não

3.5 Método de Clusterização (ou Agrupamento)

O Método de Clusterização, como o próprio nome identifica, condiz com a técnica de Mineração de Dados de Análise de Clusters. O que diferencia a Análise de Clusters de Técnicas de Classificação é a que a primeira é uma tarefa de Aprendizado Não-supervisionado, devido aos *clusters* representarem classes que não estão definidas no início do processo de aprendizagem; como é o caso das Técnicas de Classificação (Aprendizado Supervisionado), onde o banco de dados de treinamento é composto de tuplas classificadas. Clusterização constitui uma

tarefa de aprendizado por observação ao contrário da tarefa de Classificação que é um aprendizado (Amo, 2003).

O algoritmo *k-means* é um método popular da tarefa de Clusterização (Goldschmidt, 2005). Toma-se randomicamente, *k* pontos de dados (dados numéricos) como sendo os centróides (elementos centrais) dos *clusters*. Em seguida, cada ponto (ou registro de base de dados) é atribuído ao *cluster* cuja distância deste ponto em relação ao centróide de cada *cluster* é a menor dentre todas as distâncias calculadas. Um novo centróide para cada *cluster* é computado pela média dos pontos do *cluster,* caracterizando a configuração dos *clusters* para a iteração seguinte. O processo termina quando os centróides dos *clusters* param de se modificar, ou após um número limitado de iterações que tenha sido especificado pelo usuário.

O algoritmo *k-means* toma um parâmetro de entrada *k,* e divide um conjunto de *n* objetos em *k* clusters tal que a similaridade intracluster resultante seja alta, mas a similaridade intercluster seja baixa. A similaridade em um *cluster* é medida em respeito ao valor médio dos objetos neste *cluster* (centro de gravidade do *cluster*).

A execução do algoritmo *k-means* consiste em, primeiro, selecionar aleatoriamente *k* objetos, que inicialmente representam cada um a média de um *cluster.* Para cada um dos objetos remanescentes, é feita a atribuição ao *cluster* ao qual o objeto é mais similar, baseado na distância entre o objeto e a média do *cluster.* A partir de então, o algoritmo computa as novas médias para cada *cluster.* A partir de então, o algoritmo computa as novas médias para cada *cluster.* Este processo se repete até que uma condição de parada seja atingida.

O método *k-means* não é adequado para descobrir clusters com formas não convexas ou clusters de tamanhos muitos diferentes. Existem ainda os objetos que são diferentes ou inconsistentes em relação ao conjunto de dados formados (ruídos ou *outliers*) onde o método *k-means* é sensível.

Pois, o pequeno número de ruídos pode influenciar os valores médios dos *clusters*. Ainda são encontradas muitas variações do método *k-means* como, por exemplo: o *k-modes, k-prototypes* e *k-medoids*. As variações diferenciam-se na

seleção das *k* médias iniciais, no cálculo da similaridade e na estratégia para calcular a média dos *clusters.*

Um importante ponto a ser considerado é como medir o quanto um elemento é similar ao outro e verificar se pertence a um determinado *cluster* ou não. Portanto, utiliza-se a uma medida de similaridade a qual é específica para cada problema de clusterização a ser tratado (Santos, 2008). Uma das medidas (ou critérios) para medir o grau de similaridade é o cálculo da distância entre os elementos, ou seja, quanto menor a distância entre um par de elementos maior é a similaridade entre eles.

As medidas de similaridade (Hair-Jr., 2005) são utilizadas para analisar a semelhança de objetos que foram agrupados. São divididas em três: Medidas de Distância, Medidas Correlacionais e Medidas de Associação. Cada uma das três medidas representa uma perspectiva particular de similaridade, que depende de seus objetivos e da natureza dos dados. Pois, tanto as Medidas Correlacionais quanto as Medidas de Distância requerem dados métricos, ao passo que as Medidas de Associação são para dados não-métricos.

As medidas de distância representam a similaridade como a proximidade entre observações (instâncias) ao longo dos atributos. Estas medidas efetuam uma medida de dissimilaridade, em que os valores maiores denotam menor similaridade.

Como exemplos de métricas de distâncias pode-se citar: Distância Euclidiana (equação 4), Distância de Manhattan (equação 5), Distância de Minkowski (equação 6) (Amo, 2012). Abaixo segue as equações referentes a distâncias:

$$d(i,j) = \sqrt{| x_{i_1} - x_{j_1} |^2 + | x_{i_2} - x_{j_2} |^2 + \ldots + | x_{i_n} - x_{j_n} |^2} \quad (4)$$

$$d(i,j) = | x_{i_1} - x_{j_1} | + | x_{i_2} - x_{j_2} | + \ldots + | x_{i_n} - x_{j_n} | \quad (5)$$

$$d(i,j) = \sqrt[q]{(| x_{i_1} - x_{j_1} |^q + | x_{i_2} - x_{j_2} |^q + \ldots + | x_{i_p} - x_{j_p} |^q)} \quad (6)$$

Onde:

- x_i, x_j representam as características dos objetos;

- A distância de Minkowski (q > 1) generaliza tanto a distância euclidiana (caso especial onde q = 2) quanto a distância de Manhattan (caso especial onde q = 1).

As Medidas Correlacionais representam similaridades pela correspondência de padrões ao longo dos atributos Ela não olha a magnitude do valores dos atributos, apenas o padrão global de valores. Ou seja, os *clusters* baseados em Medidas Correlacionais podem não ter valores similares, mas sim padrões similares. Enquanto *clusters* baseados em distância têm valores mais similares no conjunto de atributos, mas os padrões podem ser bem diferentes. Exemplo de métrica correlacional é: a Correlação de Pearson (ou Coeficiente de Correlação de Pearson) onde segue a descrição deste algoritmo abaixo:

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\left[\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2\right]}} \quad (7)$$

Onde:

- Mede-se o nível de similaridade entre duas variáveis e x e y são os valores medidos das duas variáveis e $\overline{x}$, $\overline{y}$ são respectivamente suas médias.
- O r assume apenas valores entre -1 e 1.

A Correlação de Pearson mede a similaridade direcional entre dois pontos. O valor da correlação varia entre -1 e 1; quando for 1 significa que os dois pontos possuem exatamente o mesmo comportamento, quando for 0, que são completamente não relacionados e quando for -1 significa que são inversamente relacionados.

As Medidas de Associação são usadas para comparar objetos cujas características são apenas em termos não-métricos (medida nominal ou ordinal). Cabe ressaltar que muitas aplicações não dão todo suporte para as medidas de associação. Abaixo segue a descrição das métricas de associação Similaridade de Cosseno e Coeficiente de Jaccard respectivamente:

$$\cos(x, y) = \frac{x \cdot y}{||x||\,||y||} = \frac{\sum_i^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2}\sqrt{\sum_{i=1}^n y_i^2}} \quad (8)$$

$$s(x, y) = \frac{|x \cap y|}{|x \cup y|} \quad (9)$$

O Cálculo da similaridade usando cosseno é feito calculando-se o vetor (*cosine-based*) entre dois usuários, que são tratados como vetores (*x* e *y*) em um espaço n-dimensional, onde n é o número de itens dos vetores (Santos, 2008). A idéia é que a similaridade máxima é atingida quando o vetores apontarem na mesma direção (ângulo = 0º) e a similaridade é mínima quando o vetores forem perpendiculares (ângulo = 90º).

A medida de associação de Jaccard (Bank, 2008) expressa a similaridade entre dois conjuntos sendo definida por |*Conjunto1 interseção Conjunto2*| / |*Conjunto1 união Conjunto2*|. Ou seja, refere-se ao número de itens em comum, dividido pelo número total de itens distintos em dois conjuntos.

Após a definição de qual medida de similaridade adotada para utilização no método de clusterização é apresentado os resultados. A Figura 9 e Figura 10 ilustram a aplicação do algoritmo *k-means* em um arquivo com 20 registros de dados, considerando-se *k=3*.

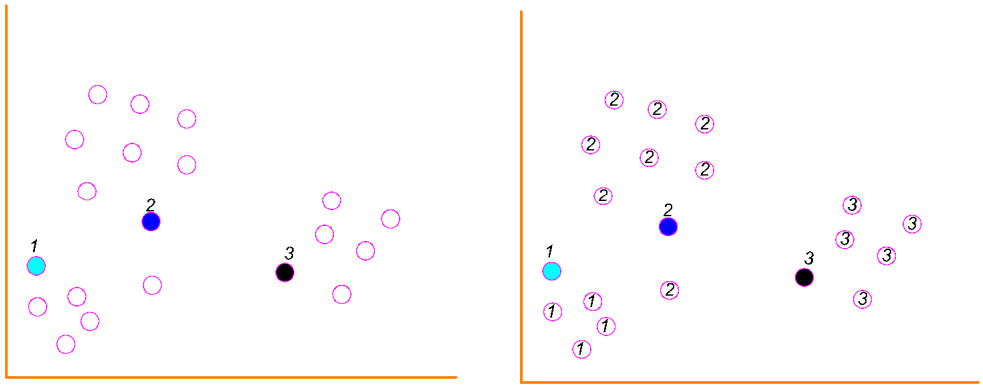

Figura 9: Algoritmo *K-Means* 1ª parte: a) Inicialização das Médias. b) Atribuição dos rótulos aos objetos.

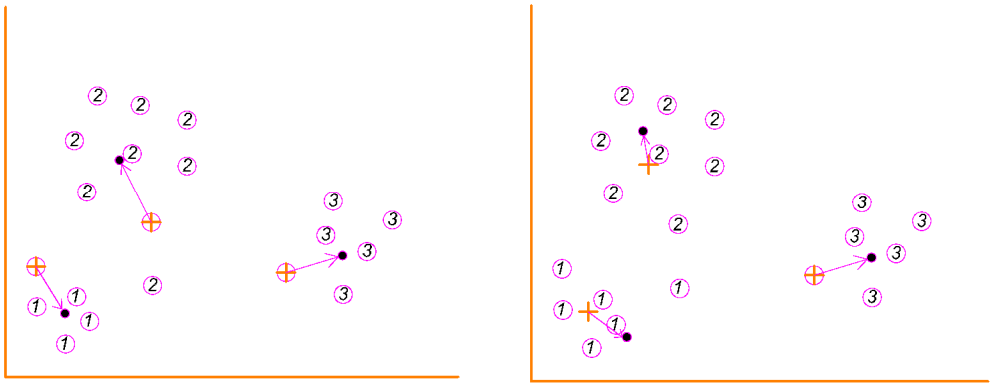

Figura 10: Algoritmo *K-Means* Passo subseqüente: a) Atualização das Médias. b) Nova atribuição de rótulos e atualização das médias.

O algoritmo *k-means* é inicializado com os centros (médias) colocados em posições aleatórias. A busca pelo centro comum se faz de forma iterativa. Após essa inicialização, os objetos restantes são agrupados conforme a distância em que se encontram das médias.

A aplicação de Mineração de Dados para a Recomendação Social proposto nesta pesquisa utiliza uma abordagem híbrida de Filtragem de Informações onde será adotado dois métodos citados anteriormente que atuem tanto em uma abordagem visando o conteúdo textual do que será filtrado quanto a colaboração existente em uma Rede Social. Portanto, conforme já descrito sobre os métodos de Mineração de Dados nas seções anteriores, os métodos ideais para a criação desta metodologia de Mineração de Dados em Texto são os métodos que trabalham com os Modelos Vetoriais e com a Clusterização; respectivamente o algoritmo TF-IDF e Similaridade de Cosseno. O primeiro algoritmo torna-se ideal para a análise dos dados textuais extraídos de uma Rede Social. O segundo algoritmo vem complementar o algoritmo TF-IDF através de uma análise voltada para a similaridade dos dados.

Posteriormente, a aplicação de Mineração de Dados apresenta uma etapa final onde são ordenados em forma de *ranking* os melhores resultados para serem recomendados ao usuário. Neste âmbito, o Método *K*-Vizinhos mais Próximos é adotado para este seguimento da pesquisa. Por apresentar um algoritmo de métrica simples e com todas as condições de agrupar os resultados já manipulados pelos algoritmos da abordagem híbrida.

4 REDE SOCIAL

Atualmente, os relacionamentos de pessoa a pessoa tem se manifestado fortemente em ambientes tecnológicos através das Redes Sociais. O relacionamento se estende desde as pessoas não tanto conhecidas na sociedade quanto as pessoas famosas (celebridades, esportistas, autoridades públicas e etc.); através de um vínculo em comum ou não, ou ainda por agregar características existentes nos perfis das pessoas, que se traduzem em comunidades ou grupos de pessoas de interesse similar.

4.1 Contexto Histórico

A reunião de mais de uma pessoa em torno de: um respectivo tema, *hobby*, ideologia, ou pelo mesmo interesse na compra de determinado produto, ou ainda interesse em lugares a se freqüentar existia na Internet antes mesmo da definição do termo “Rede Social”. Inicialmente é correto destacar que o estudo das Redes Sociais se apresenta em vários campos: filosofia, educação, psicologia, ciência e nos últimos anos tem se delimitado a computação (sobretudo no campo de pesquisa da Inteligência Artificial).

No passado, o estudo das Redes sociais era feito apenas no campo da sociologia e antropologia através de ferramentas típicas adotadas em entrevistas e questionários (Wasserman et al., 1994). O que acontecia eram pesquisas realizadas com pequenas bases de dados e pouco representativas. Com o surgimento de redes sociais consolidadas como *Orkut*, *Facebook*, *Twitter* e similares; é que surgiu a oportunidade de estudos sobre redes sociais com o uso de grandes bases de dados. A Tabela 3 apresenta a relação de algumas das Redes Sociais mais pertinentes e suas caracterizações.

Tabela 3: Quadro-comparativo com algumas das Redes Sociais conhecidas.

Nome	***Facebook***	***Friendster***	***Bebo***	***LinkedIn***	***MySpace***	***...VZ***	***Twitter***	***Xing***	***TheNext***	***Foursquare***
Tipo	Livre	Livre	Mídias	Profissional	Mídia	Livre	Livre	Profissional	Varejista	Geolocalização
Fundação	2004	2002	2005	2003	2003	2005	2006	2003	2006	2009
Grupo Alvo	Todos do Mundo	Todos do Mundo	Todos (entusiastas por vídeos e música)	Especialistas e profissionais.	Todos do Mundo	Estudantes de todo o Mundo	Todo o Mundo	Apenas profisionais	Todo o Mundo.	Todo o Mundo
Propósito	Amizade	Amizade	Amizade	Profissional	Amizade	Estudante	Amizade	Profissional	Mercado	Amizade
Acesso	*Web, apps*	*Web, apps*	*Web*	*Web, apps*	*Web*	*Web*	*Web, mobile*	*Web, apps*	*Web*	*Web, móbile, apps*
Membros	500 milhões	115 milhões	10.7 milhões	75 milhões	270 milhões	17 milhões	Mais que 75 milhões	_	Mais de 600 milhões.	Mais de 10 milhões
Avaliados	700-800 milhões de dólares	_	_	Mais 17 milhões de euros.	495 milhões de dólares.	18 milhões de euros.	0, capital financi-ado	45 milhões de euros.	_	500 milhões de dólares

Entretanto, a teoria moderna das Redes Sociais teve um início bem mais cedo no ano de 1967 com as pesquisas de Stanley Milgram (Kumar, 2002). Seu trabalho consistia em realizar experimentos onde haveria uma comunicação por cartas entre diversas pessoas residentes em Omaha e Nebraska para pessoas que residiam em Boston. A lógica seria que as pessoas só poderiam enviar a carta para outra pessoa que elas conhecessem pelo primeiro nome e a resposta das endereçadas também seria para o primeiro nome. O objetivo era de que a carta chegasse ao seu associado no menor número de "passos" possíveis. A estes "passos", Milgram chegou a conclusão que o número médio dos passos serviria para descobrir o número de sucesso das cartas que conseguiam chegar ao seu local, sendo seis este numero ideal. Ou seja, quaisquer duas pessoas residentes nos Estados Unidos estavam ligadas em uma Rede Social com "seis graus de separação".

Uma definição simples com respeito à Rede Social é que se trata de um grupo de pessoas que mantém relações e interações em grupo. Fazendo um comparativo com terminologias técnicas: as pessoas seriam como "nós" em um gráfico e as relações entre as pessoas como "*links*" ou "*tags*".

Outra definição seria que as Redes Sociais são um conjunto de links que organizam as pessoas, grupos e instituições de forma igualitária e democrática, e em torno de um objetivo comum (Feldman-Bianco, 1987).

O pesquisador Benevenuto (2010) descreve de forma mais restrita que a Rede Social é um serviço Web que permite que indivíduos construam perfis públicos ou semi-públicos dentro de um sistema, com a proposta de articular uma lista de outros usuários com os quais compartilhará (ou não) conexões; e visualizará percorrerendo suas listas de conexões e outras listas feitas por outras pessoas no sistema.

O conceito de Rede Social está também intimamente ligado ao de "Mídias Sociais", pois adota modelos de comunicação de forma descentralizada e independente de controles editoriais de algum grupo. Segundo Machado (2010), esta não é uma simples tendência, e sim uma nova forma de comunicação, que vai à contramão do que as organizações geralmente vêm

fazendo, de cima para baixo, ou seja, comunicações sem nenhuma interação ou diálogo com o “público” gerando a democratização do diálogo.

Ou seja, a Rede Social é um modelo dinâmico e flexível, com liberdade e espontaneidade entre os elos, o respeito pela individualidade dos usuários sendo baseado principalmente na confiança mútua entre os usuários que dela fazem parte. Os membros de uma Rede Social podem recolher e divulgar dados, informações e conhecimento. Ficando a critério de cada um aceitar, ou não, estas informações disponibilizadas. Alguns autores chamam a atenção para a intensidade em que estes elos (ou laços) são formados. Pois, dependendo da força que existe entre as relações é que se caracterizarão os diferentes tipos de laços. A este tipo de laço convém destacar que: uma ligação forte é aquela estabelecida diretamente entre duas pessoas na mesma Rede Social, enquanto uma ligação fraca é uma relação entre duas pessoas conectadas através de outra pessoa (dois níveis de separação).

Segundo análise feita por Jamali (2006), a Rede Social pode ser entendida por uma espécie de gráfico que mostra a sua composição através de pontos (ou nós) para representar os atores e as linhas (ou bordas) para representar laços ou relações. O pesquisador destaca que os sociólogos pensaram esta maneira de representar graficamente a partir dos matemáticos, eles mudaram os seus gráficos para "sociogramas". Há uma série de variações sobre o tema de sociogramas, mas todos eles compartilham a característica comum de usar um círculo marcado para cada ator na população (grupo) que fazem parte, descrevendo por segmentos de linha entre os pares de atores para representarem que existe um laço entre os dois. É enfatizado que este gráfico apresenta uma abordagem simples para poder representar toda complexidade de uma Rede Social. Abaixo segue a Figura 11 com a proposta de Jamali (2006) e a Figura12 que ilustra a idéia de grafos representando as relações dos perfis em nós.

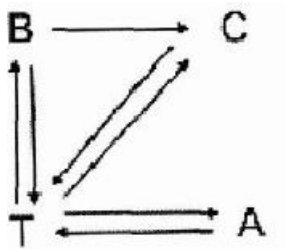

Figura 11: Representação da relação em nós da Rede Social (Fonte: Abolhassani, 2006)

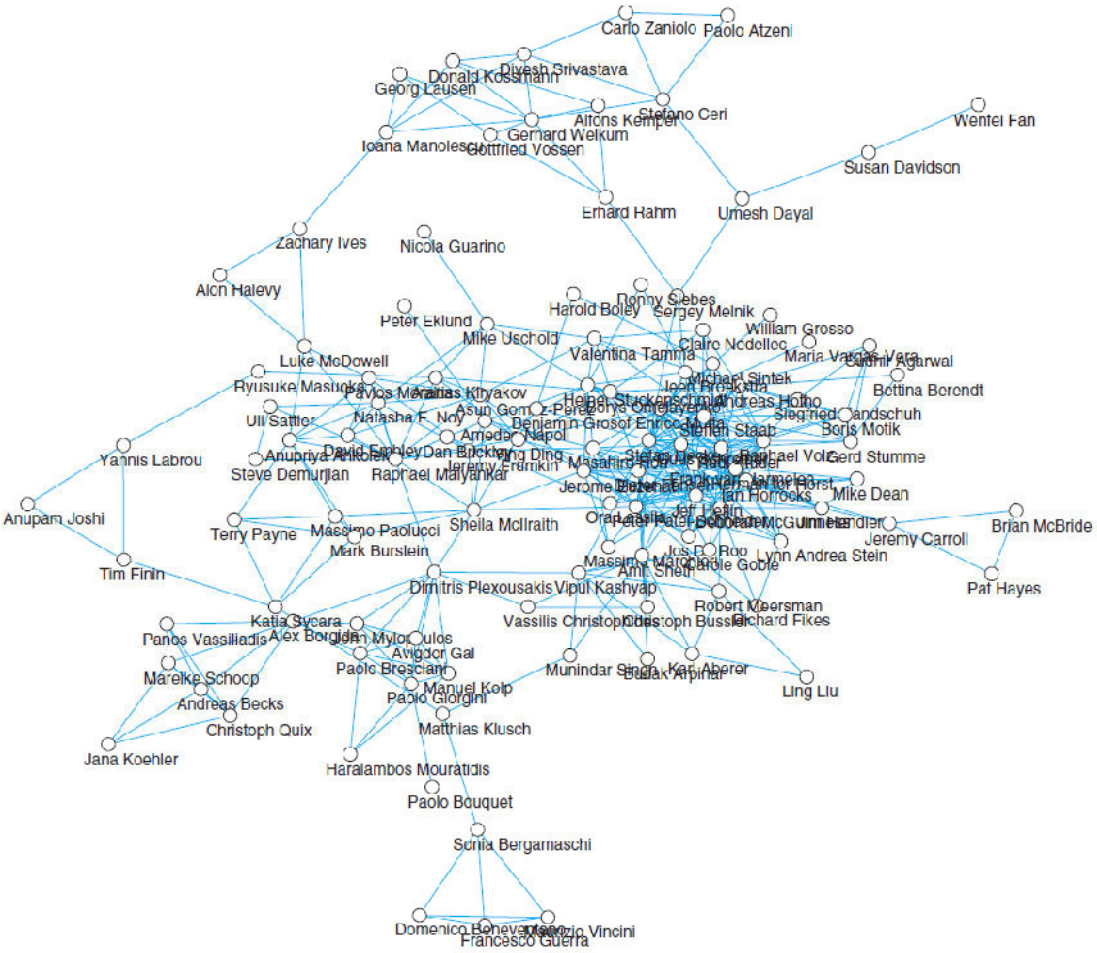

Figura 12: Estrutura de um Grafo em uma Rede Social.

Quanto as aplicações possíveis, as Redes Sociais, seguem vários campos onde se pode destacar com maior participação:

- No campo empresarial, por exemplo, poderia ter uma situação de análise dos trabalhadores organizados em redes sociais; com intuito de

reter o conhecimento (memória corporativa) com o intuito de evitar os problemas típicos que dificultam a difusão de conhecimento dentro do ambiente de serviço e nas tomadas de decisão dentro da empresa.

- No campo da ciência, as redes sociais podem ajudar a estudar a propagação de endemias ou epidemias, através de geolocalização destas em mapas temáticos. Contribuindo para o entendimento do aumento e diminuição da proliferação de doenças.

Na profusão de propaganda e marketing, a Rede Social poderia ser utilizada como ferramenta para ajudar uma marca, a venda de um produto, serviços e similares. Com o intuito de fazer sua disseminação em grupo específico da área: patrocínio de conteúdos interativos; pesquisa de idéias para novos produtos; promoção de produtos e serviços em redes sociais online.

Segundo o pesquisador Liccardi (2007) os tipos de algoritmos investigados para a concepção de uma Rede Social podem ser descritos como:

- ***Searching networks*:** Em busca de critérios específicos dentro de uma Rede Social, Zhang e Ackerman (2005) estudaram as características sociais de vários algoritmos de pesquisa que podem ser úteis para detectar características individuais, tais como perícia, a fim de compreender as vantagens e desvantagens envolvidas no design de motores sociais baseados em rede de buscas. O uso de algoritmos de busca para navegar nas Redes Sociais pode ser altamente benéfico na procura de uma pessoa especial e depois identificar as pessoas ligadas a ela.
- ***Constructing networks*:** A maneira com que as pessoas se encontram e a forma que as Redes Sociais agem na vida cotidiana das pessoas chamam a atenção de pesquisadores de ciências da computação. O fato de como se relacionam e dependem da Rede Social para algumas atividades como a amizade, o apoio, interesses especiais e partilha de conhecimento inspirou aos desenvolvedores um algoritmo para analisar essas facetas de uma forma mais abstrata. Hamasaki e Takeda (2003) propõem uma metodologia *Matchmaker* vizinha, onde dois indivíduos que não se conhecem são introduzidos por um mediador, este é um

amigo de ambos, e que pode facilitar a criação de um novo relacionamento. Um método semelhante de criação ou para ampliar uma Rede Social seria seguir o amigo de um amigo, onde as associações são inferidas através de amigos mútuos. No entanto, com métodos como o amigo de um amigo, problemas de confiança e privacidade se tornam questionáveis devido à discutível e inferência de medição de confiança. Redes Sociais podem servir como uma rica fonte de novo conhecimento e como um filtro para identificar a informação mais pertinente para as necessidades específicas do usuário.

- ***Network dynamics*:** Alguns pesquisadores ressaltam o quanto pode ser diferente a compreensão de como as redes sociais desempenham papel na formação de comunidades. O pesquisador Wellman (2001) vê a diferença entre as redes e grupos: “Embora as pessoas vejam o mundo em termos de grupos que funcionam em redes. Em sociedades em rede, fronteiras são permeáveis, interações com os outros são diversas, conexões alternam entre redes e hierarquias, por ser mais lisa e recursiva”. Devido à transparência e natureza vaga das redes sociais, os membros movem-se dentro e fora das comunidades, sem formalismo. Ou seja, a pessoa ao invés de se encaixar no mesmo grupo que a rodeia pode dar preferência a uma comunidade com que se identifique. Inspirado nestas interações dentro das Redes Sociais, os pesquisadores usam modelos matemáticos para simular o que acontece na vida diária. Os exemplos de algoritmos, neste contexto, incluem: a formação de coalizão, a formação de redes e de estabilidade, algoritmos de agrupamento (*clusters*), algoritmos de clubes, e algoritmos com teoria dos jogos. Aplicações destes algoritmos ocorrem em uma variedade de campos que inclui processamentos distribuídos, comunicação e redes de computadores, a economia social, e jogos *multiplayer*. Outro exemplo seria no campo da educação: consultas ou visualização nas redes sociais de estudantes em diferentes áreas para fins de avaliação. Por exemplo, se o professor tem uma visualização da rede, esta pode facilmente reconhecer casos de plágio dado que os amigos dos alunos e seus colegas são mostrados na rede.

4.2 Elementos e Formação de uma Rede Social

O pesquisador Boyd (2007) destacou que existe uma estreita relação entre a identidade do indivíduo e seu perfil dentro da Rede Social. Portanto, este enumerou alguns tópicos que são comuns a todas as redes sociais e ajudam na manutenção do usuário e seu perfil:

- **Perfis**: a Rede Social geralmente traz uma página do perfil do usuário com a descrição deste membro. A idéia é que o perfil não somente identifique um indivíduo em um sistema de Rede Social, mas, que também identifique os "gostos" e hábitos em comum entre um ou mais perfis através das suas descrições. Nos perfis existem: detalhes demográficos, localização, idade, sexo, interesses (passatempos, bandas favoritas, etc.), e uma foto. Além de uma descrição de texto, imagens e outros objetos criados pelo usuário, o perfil na Rede Social também contém mensagens de outros membros e listas das pessoas identificadas como amigos na Rede Social.
- **Comentários**: grande parte das redes sociais permite que os usuários comentem e compartilhem seus comentários, ou ainda comente em outros perfis de usuários. Os comentários servem como base para estabelecer a comunicação em redes sociais online. Na Rede Social *Youtube* os vídeos recebem comentários, no *Facebook* e *Flickr* são as fotos que recebem os comentários, usuários do *Live Journal* podem postar seus comentários em *blogs*, etc.
- **Atualizações**: são formas efetivas de ajudar os usuários a descobrirem conteúdos e serve de encorajamento para que estes compartilhem o seu conteúdo e acessem os conteúdos dos amigos, ou amigos de amigos (geralmente estas atualizações ficam visíveis aos amigos na Rede Social). Estas atualizações contribuem no processo de aquisição de novos usuários.
- **Favoritos:** é padrão nas redes sociais que existam listagens de favoritos do conteúdo respectivo ao usuário (comidas, músicas, vídeos, *hobbys*, etc.). Isso é um fator que contribui no gerenciamento do conteúdo de

cada perfil e também como uma possível recomendação social, pois, os outros usuários podem ter acesso a estas listas (por exemplo, o *Foursquare*, *Flickr* e *Youtube*, entre outros).

- **Metadados:** contribuem na recuperação de informações em redes sociais através da marcação de conteúdo (pelos usuários): títulos, descrição, *tags* ao conteúdo que se deseja compartilhar.
- **Rankings (Lista de Top):** os rankings do tipo de conteúdo que é apresentado nas redes sociais diferenciam (através de listas) os conteúdos mais populares e menos populares. As listas são elaboradas através das avaliações usuários ou estatisticamente pelo próprio sistema da Rede Social.
- **Avaliações:** são descritas nas redes sociais no que diz respeito ao conteúdo compartilhado por um usuário. A avaliação pode ser feita por outro usuário. Por exemplo, no *Facebook* os usuários avaliam algum comentário ou conteúdo compartilhado através do termo “curtir”, no *Foursquare* o termo adotado seria “curtir esta dica”, e assim sucessivamente para outras redes sociais. Avaliações têm muito a ajudar na recomendação social, pois, por exemplo, permitem diferenciar o grau da utilidade de determinado conteúdo e sua relevância.

4.3 Rede Social Baseada em Localização

A alta distribuição de dados de localização em informações publicamente disponíveis tem acontecido em redes sociais como *o Facebook, Google, Twitter, Flickr* e *Foursquare* e gerado o que chamamos de Redes Sociais Baseadas em Localização – LBSN’s (*Location-Based Social Network*) ou Geolocalização.

Os serviços que são oferecidos neste ambiente permitem aos usuários compartilhar informações em forma de texto, imagens e vídeos – embora a disponibilidade das informações por medida de segurança seja controlada, há um enorme número de usuários que utilizam em suas informações pessoais

dados geolocalizados, transparecendo os locais que determinado usuário visitou ou que ainda está presente.

Um conceito chamado "*geotagging*" permite ao usuário que os dados de localização sejam publicados como metadados juntamente com a informação a ser compartilhada. Os serviços de mídia nas Redes Sociais fazem uso de dados de localização em imagens, vídeos e utilizam um sistema operacional com técnicas de *geotagging* a ser aplicado no conteúdo enviado pelo usuário. Este dinamismo serviu para o crescimento dos dispositivos móveis (celulares, *smartphones* e outros similares), pois a utilização destes recursos, pelos usuários que fornecem seus dados, não requerem o uso de máquinas de maior desempenho para alcançar o resultado esperado, e sim apenas a conexão com a Internet e tecnologia GPS.

Entre as LBSN's consolidadas pode-se citar: *Google Places, Foursquare, MeetMoi, Google Latitude, Brightkite* e *Gowalla*. Estas Redes Sociais tem como característica permitir que usuários verifiquem sua localização atual, ou encontrem a localização da sua empresa, ou ainda encontre pessoas próximas (ou não) que transmitam algum grau de interesse ao usuário. O objetivo é que a partir da localização encontrada, possa se compartilhá-la em um ambiente dinâmico, e através de comentários (informações textuais) manifestarem opiniões (recomendações) a respeito dos determinados locais freqüentados ou das pessoas que ali se encontram. Este registro da localidade apresenta um nome comum em grande parte das LBSN's que trabalham com dados geolocalizados conhecido pelo termo: *"checkin".*

Este Trabalho de Dissertação adota o *Foursquare* como a Rede Social com o ambiente propício para criação da aplicação de Mineração de Dados, por ser, atualmente, uma das maiores redes sociais deste âmbito (até o momento da escrita deste trabalho, aproximadamente mais de 10 milhões de usuários) e apresentar excessiva abordagem colaborativa em suas relações sociais, que necessita de técnicas de abordagem de conteúdo mais efetivas para gerar resultados expressivos na recomendação social. O Capítulo 6 onde trata a criação da metodologia desenvolvida para a pesquisa da Dissertação apresentará mais detalhes sobre as LBSN's e os trabalhos relacionados.

5. SISTEMAS DE RECOMENDAÇÃO

Devido à grande quantidade de informação disponibilizada pela Internet, há um conjunto de opções de rica diversidade a ser utilizada por aqueles que a acessam. Mesmo que o usuário tenha pouca experiência no assunto, a dificuldade existe em quais escolhas devem ser feitas referente ao critério que o usuário busca. Uma alternativa seria confiarmos nas recomendações feitas pelas pessoas que conhecemos, ou ainda, por outros veículos como: textos, opiniões de terceiros (revisores de filmes e livros), impressos de jornais, mídias e afins. O Sistema de Recomendação entraria como um mediador para as necessidades auxiliando o processo de indicação através da interação social junto a um ou mais indivíduos.

No ambiente *Web*, os Sistemas de Recomendação são conhecidos pelo papel de “conselheiros” dos usuários; ajudando-os na escolha de determinado item, seja na forma de adquirir este item ou apenas examiná-lo. Segundo Schafer (1999), os Sistemas de Recomendação são utilizados para identificar usuários, armazenar suas preferências e recomendar itens que podem ser produtos, serviços e/ou conteúdos, de acordo com suas necessidades e interesses.

Para atingir este objetivo é necessária a formulação de estratégias para a recomendação e adotar uso de técnicas que prezem tanto pelo conteúdo dos itens que se deseja recomendar assim como o possível relacionamento entre usuário e item.

5.1 Contextualização e Estratégias de Recomendação

Uma definição para os Sistemas de Recomendação é que auxiliam no aumento da capacidade e eficácia do processo de indicação, já bastante conhecido na relação social entre os seres humanos (Resnick e Varian, 1997).

O autor Schafer (2001) destaca que os Sistemas de Recomendação são utilizados pelos sites de comércio eletrônico com intuito de sugerir produtos para seus clientes e fornecer informações que procuram auxiliar os clientes sobre qual produto deve ser adquirido.

Alguns autores destacam que os Sistemas de Recomendação apresentam a capacidade de identificar e aprender as preferências e necessidades de um usuário, gerando recomendações customizadas ao seu perfil – estes recebem o nome de Sistemas de Recomendação Personalizados (Schafer, 2001) e (Teixeira, 2002).

A principal preocupação dos Sistemas de Recomendação é o tratamento das informações que são produzidas e disponibilizadas no ambiente dinâmico da Internet, pois a enorme quantidade de informações poderá ocasionar uma sobrecarga aos usuários. Portanto, a seleção e filtragem das informações para recomendar ao usuário é de suma importância tendo sido discutido há alguns anos por outros pesquisadores, como Peter Denning (1982), Goldberg (et al. 1992), Resnick (et al. 1994) e Shardanand & Maes (1995).

O primeiro Sistema de Recomendação denominado *Tapestry* (Goldberg et al., 1992) e (Resnick & Varian, 1997) define um tipo de sistema específico no qual a filtragem de informação era realizada através de auxilio humano, ou seja, a colaboração entre os grupos de pessoas interessados na recomendação. Os usuários de *newsgroups* podiam selecionar outros usuários de quem gostariam de ouvir, devido a essa característica recebeu como nome a expressão “Filtragem Colaborativa”.

Este aspecto colaborativo gera uma Recomendação Social onde retorna hábitos pessoais, produtos, serviços ou recomendação de outros usuários. Esta implicação social além de reunir as recomendações por um conjunto de usuários ou serviços requisitados em comum similaridade; contribui para formular agregações personalizadas em que a grande variedade de “gostos” diferenciados enriquecerá a recomendação.

A estrutura de um Sistema de Recomendação é dividida em quatro processos: identificação do usuário, coleta de informações, estratégias de recomendação e visualização das recomendações. Abaixo, segue a Figura 13 ilustrando a estrutura básica de um Sistema de Recomendação.

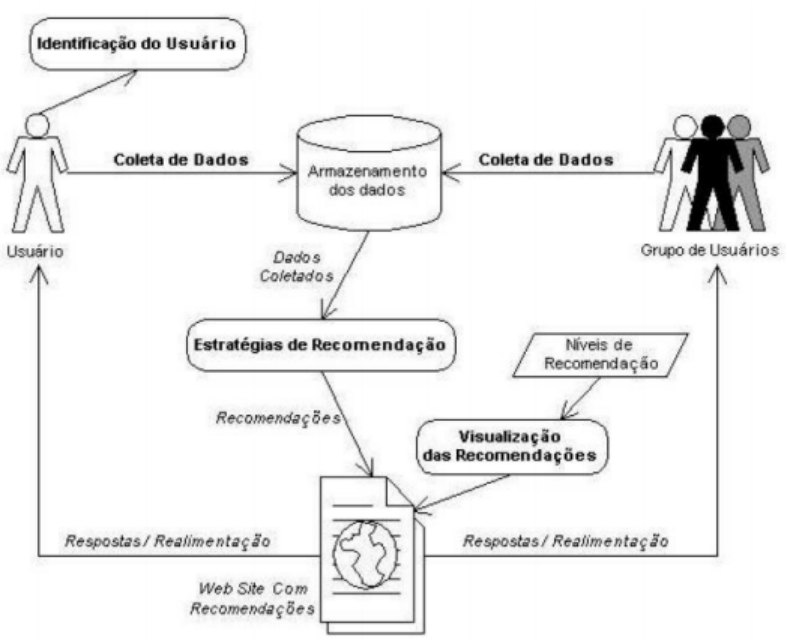

Figura 13: Estrutura de um Sistema de Recomendação (Schafer et. al., 2000).

Atualmente os Sistemas de Recomendação se caracterizam pela personalização como fator incisivo ao usuário; identificando seus "gostos" e características pessoais para representação em um chamado "perfil de usuário". A princípio a primeira fase da estrutura dos sistemas é a coleta de dados do usuário, seja este apena um único perfil ou um grupo de usuários com perfis diferenciados, e, por conseguinte armazenar os dados em um repositório.

Esta coleta de dados do usuário simboliza uma proposta de modelagem do comportamento do usuário. Ou seja, a prática da recomendação é realizada através do conhecimento a respeito de dados pessoais do usuário e a forma que ele acessa o sistema. Um bom exemplo são os sites de comércio

eletrônico, onde o comportamento do usuário no sistema caracterizará os produtos que lhe interessam em detrimento de outros de menor interesse, ou ainda a identificação dos usuários com hábitos semelhantes ou diferenciados.

Inicialmente, a identificação do usuário segue como aspecto primordial para a coleta dos dados corretamente. Segundo Reategui & Cazella (2005), antes que se possa recomendar algo aos usuários é necessário que se reconheça os seus comportamentos habituais e relativos e, portanto, a identificação acontece de dois modos: identificação no servidor (cadastro com as informações pessoais do usuário para ser armazenado em um banco de dados no servidor de forma que o sistema reconheça o usuário que acessa) e identificação no cliente (utiliza o recurso dos *cookies* onde o *website* consegue identificar que aquele computador se conecta a ele com freqüência ou não, identificando assim seu usuário).

Para adquirir os dados do usuário e modelar seu comportamento a coleta dos dados pode ser feita em dois âmbitos: uma aquisição explícita e uma aquisição implícita (Claypool et al., 2001). Na primeira opção o usuário apresenta o seu grau de interesse em um determinado item através de uma nota, por exemplo, construindo uma escala quantitativa para ajudar na recomendação, o usuário pode descrever espontaneamente que existem duas seções que lhe são favoritas em um *website*, por exemplo. A aquisição implícita reúne algumas ações do usuário no sistema – em um *website,* por exemplo – que possam de alguma forma ser utilizadas para inferir em suas preferências. Entretanto o problema crítico está na importância dos critérios a serem adotados para formação do perfil do usuário (Bezerra, 2004).

Alguns pesquisadores se dedicaram a estudar a respeito das práticas de aquisição de dados dos usuários pelos Sistemas de Recomendação. Seja através de comportamentos simples no sistema, como: salvar, imprimir ou adicionar itens favoritos (Oard e Kim, 1998) e (Chan, 1999); quanto à análise de comportamentos de maior complexidade: o tempo gasto de leitura de cada usuário em determinadas pagina *web* (Morita e Shinoda, 1994), (Chan, 1999) e (Konstan et al., 1997); e ainda adquirir mais informações pela forma que o

usuário manuseia o navegador *web* (*web browser*), capturando seleções de texto ou *clicks* do mouse Goecks e Shavlik (2000).

Segundo a estrutura ilustrada na Figura 13, os dados coletados serão utilizados em estratégias visando à recomendação aos usuários, de acordo com a especificidade do sistema. Os autores Reategui & Cazella (2005) apresentam um estudo sobre as estratégias de recomendação listadas nas seguintes opções:

- **Listas de Recomendação**: consiste em manter listas de itens organizados por tipos de interesses. Não havendo a necessidade de uma análise mais profunda de dados do usuário e sim apenas a observação dos tipos de itens mais populares, e ordenação destes em grupos que o destaquem, por exemplo: "Itens mais vendidos", "Idéias para presentes", entre outros.

- **Avaliações de Usuários**: Uma das estratégias mais utilizadas em Sistemas de Recomendação são as avaliações dos usuários. Ou seja, em um *website* de comércio eletrônico além de comprar um produto o usuário também deixa um comentário sobre o item adquirido, de modo que sirva para assegurar a qualidade de determinado item.

- **As recomendações do usuário**: de recomendação é oferecido em uma seção inteiramente dedicada a sugestões feitas especificamente para o usuário. Onde dois tipos de recomendação são possíveis nestas seções: aquelas feitas a partir de preferências implícitas (elementos moldados através do comportamento do usuário) ou explícitas (elementos adquiridos pelo perfil do usuário).

- **Associações entre os "gostos" dos usuários**: esta recomendação acontece quando é detectada uma associação entre itens avaliados por usuários; sendo considerada umas das estratégias mais complexas.

- **Associação por conteúdo**: Também é possível fazer recomendações com base no conteúdo de determinado item. Por exemplo, um autor, um compositor ou um editor de determinado livro.

No próximo tópico serão apresentadas as técnicas que os Sistemas de Recomendação utilizam para efetuar as estratégias acima citadas e que atuam na implementação desta pesquisa de Dissertação.

5.2 Técnicas de Filtragem de Informação em Sistemas de Recomendação

Anteriormente foram apresentadas as estratégias de recomendação que são trabalhadas em vários Sistemas de Recomendação. Para a utilização destas é necessária a utilização de algumas abordagens de técnicas de Filtragem das Informações, tais como: Filtragem de Informação Baseado em Conteúdo, Filtragem de Informação Baseada na Colaboração e Filtragem de Informação Híbrida.

Para os autores Belvin e Croft (1992), a Filtragem de Informação é o nome utilizado para descrever uma variedade de processos que envolvem a entrega de informação para as pessoas que realmente necessitam delas. As técnicas de Mineração de Dados citadas no capítulo 2 e seus algoritmos citados no Capítulo 3 são utilizadas nas abordagens de Filtragem de Informação, que ao final do processo, resultam na recomendação desejada.

5.2.1 Filtragem de Informação Baseada em Conteúdo

Os autores Balabanovic e Shoham (1997) descrevem que os Sistemas de Recomendação que adotam este tipo de técnica têm como objetivo gerar de forma automática a descrição dos itens e comparar com a descrição de interesses do usuário ou seu histórico de “navegação”, de forma que possa sugerir o que for mais relevante ao usuário. Devido a filtragem poder realizar uma seleção baseada na análise de conteúdo dos itens e no perfil do usuário recebe o nome de Filtragem de Informação Baseada em Conteúdo (Herlocker, 2000). A preferência do usuário pelos itens freqüentemente é usada para

construir um perfil contendo seus indicadores de interesse baseado em determinados tópicos, geralmente representados através de um conjunto de palavras-chaves junto a pesos que possam identificar à relevância de cada item.

Segundo o autor Bezerra (2004), a aquisição das preferências do usuário nesse tipo de filtragem depende em especial da descrição dos itens que ele avalia. O ideal é que a partir do perfil coletado, seja possível a recomendação de itens mais similares aos itens bem avaliados pelo usuário e ao mesmo tempo mais dissimilares dos itens mal avaliados. A descrição de interesses do usuário é obtida através de informações fornecidas por ele próprio ou através de ações, como seleção e aquisição de itens. Portanto, a preferência do usuário é freqüentemente utilizada na construção de um perfil que contenha indicadores do interesse do usuário sobre determinados tópicos. Ou seja, a descrição dos interesses do usuário é obtida através de uma consulta, ou aprendendo com os itens que o usuário consome – aqueles que o usuário gostou (Cazella, 2006). Abaixo segue a Figura 14 que ilustra uma avaliação feita por usuários aos itens de interesse que condizem à recomendação de filmes (Bezerra, 2004).

Filme	Gênero	Ano	País	Elenco	Diretor	Sinopse	Nota
M_1	G_1	Y_1	C_1, C_4	A_1, A_2, A_3, A_7	D_3, D_4	Bla bla ...	* *
⋮							
M_p	G_2	Y_3	C_4	A_1, A_4, A_5, A_6	D_3	Bla bla bla ...	* * * * *

Figura 14: Representação dos itens de cinema avaliados pelos usuários (Bezerra, 2004).

Muitas ferramentas que trabalham com esta abordagem aplicam técnicas como indexação de freqüência de termos (Cazella, 2006 apud Herlocker, 2000) para caracterizar o conteúdo dos usuários.

Além da indexação de freqüências de termos citadas anteriormente existem outros recursos que foram explorados por alguns autores, como por

exemplo: índices de busca booleana, onde a consulta constitui-se em um conjunto de palavras-chave que são unidas por operadores booleanos (Cazella, 2006 apud Herlocker, 2000); raciocínio probabilístico é aplicado para determinar a probabilidade que um documento tem de atender as necessidades de informação de um usuário, é conhecido pelo nome de Sistemas de Filtragem probabilística uso de linguagem natural, para criar interfaces consultas em sentenças naturais (Herlocker, 2000).

No que diz respeito aos algoritmos de Mineração de Dados a serem utilizados podemos destacar alguns exemplos: o algoritmo do vizinho mais próximo conhecido por *KNN* (*K-Nearest Neighbor*) (Cover, 1974), e outros algoritmos de aprendizagem baseada em instância (Aha et al., 1991) como algoritmos suscetíveis para serem adotados na proposta dada aos Sistemas de Recomendação. Outros exemplos são: algoritmos como o TF-IDF (Santos, 2008) e *K-D Trees* (Bentley, 1975) ajudam na eficiência da recomendação com a criação de pesos para os itens que já foram avaliados pelos usuários e criação de uma árvore binária que organiza os itens avaliados.

Entretanto, uma das principais desvantagens está nas limitadas formas de representação do conteúdo dos itens. Em alguns casos os tipos de informações filtrados não podem ser representados de forma satisfatória usando apenas variáveis de escala quantitativa ou qualitativa.

Reategui & Cazella (2005) apontam mais algumas deficiências da abordagem baseada em conteúdo. Quando o conteúdo de dados é pouco estruturado a análise é difícil (por exemplo, vídeo e som) e o entendimento do conteúdo do texto pode ser prejudicado devido a uso de sinônimos; outro problema pode ocorrer através da super especialização, pois o sistema procura se basear em avaliações positivas e negativas feitas pelo usuário, deste modo não apresentam conteúdos que não fechem com o perfil. Outra situação que torna difícil a análise dos dados é a chamada “superespecialização” dos itens de interesse do usuário. O Sistema de Recomendação apresenta uma deficiência em recomendar novos itens ao usuário, pois o sistema segue um padrão de itens a serem recomendados de acordo com uma avaliação dos tópicos de interesses dos usuários.

Este tipo de filtragem leva em muita consideração o aspecto avaliativo dos itens pelo usuário, no entanto, a quantidade de itens avaliados pelo usuário pode aumentar consideravelmente o tempo de classificação e a memória utilizada pelo sistema.

5.2.2 Filtragem de Informação Baseada na Colaboração

Segundo Herlocker (2000), a Filtragem de Informação Baseada na Colaboração (ou Filtragem Colaborativa) veio suprir algumas deficiências que ficaram com a utilização da Filtragem Baseada em Conteúdo. Pois o foco desta técnica é a colaboração entre as pessoas e suas experiências em vários seguimentos e não fixa interesse específico na compreensão e conhecimento dos itens.

O primeiro Sistema de Recomendação *Typestry* (Goldberg et al., 1992), comentado no início do capítulo, era conhecido por utilizar esta abordagem, quando o usuário especificava uma consulta como: “mostre-me todos os memorandos que uma determinada pessoa considera como importante” (Cazella, 2006), desta forma outros usuários poderiam ter acesso ao tópico de interesses de outro usuário.

Os Sistemas de Recomendação que iniciaram com a Filtragem Colaborativa exigiam de forma explícita uma predição entre as opiniões dos usuários de forma que identificassem os itens de interesse e o possível grau de relação. Com o tempo este processo foi automatizado através de pontuações aos itens de interesse, sem haver a necessidade de se requerer ao usuário deste artifício. Alguns automatizaram as relações entre usuários através de algoritmos de vizinhos mais próximos, descoberta de padrões comuns de comportamento, ou ainda formação de grupos e comunidades que compartilham de mesmo interesse em determinados itens.

Um exemplo de ambiente baseado em Filtragem Colaborativa é o Sistema de Recomendação de filmes *MovieLens* (Riedl et al., 1999). O usuário insere pontuações para filmes que tenha visto e o sistema utiliza estas pontuações para encontrar pessoas com gostos similares. O sistema

recomendar filmes nos quais indivíduos com gostos semelhantes se interessaram, mas que não assistiram ainda.

A colaboração denota vantagens na possibilidade de apresentar aos usuários recomendações inesperadas (uma deficiência da Filtragem Baseada em Conteúdo). Ou seja, o usuário receberia recomendações de itens que não estavam seguindo o padrão preestabelecido em um perfil de usuário.

A técnica de Filtragem Colaborativa baseada na abordagem do “vizinho mais próximo” (*KNN*) permite a vantagem de rapidamente incorporar na lista de recomendações de um usuário itens totalmente novos (Bezerra, 2004). Entretanto, em grupos de usuário de maior contingente torna-se complexo, devido a lentidão da busca pelos usuários mais semelhantes ao usuário alvo.

A Filtragem Colaborativa também apresenta suas desvantagens. No momento em que um novo item é adicionado ao sistema, o momento de sua primeira recomendação pode levar algum tempo, pois este precisa ser bem avaliado por um número significativo de membros da comunidade. Pois esta técnica de filtragem depende da experiência dos membros da comunidade para determinar a relevância de um item. Se a base de informação cresce ou atualiza-se rapidamente ou ainda é muito maior do que o número de usuários, a qualidade das recomendações de um sistema baseado em Filtragem Colaborativa pode ser comprometida (Bezerra, 2004).

Outra desvantagem nesta técnica de filtragem é se o sistema comporta perfis dos mais variados entre os membros de uma comunidade. A similaridade entre os perfis torna-se complexa para que o sistema gere uma recomendação eficiente.

Assim como na Filtragem Baseada em Conteúdo, algoritmos de Mineração de Dados podem ser incorporados como forma de suprir as deficiências da Filtragem Colaborativa, onde merecem destaque: redes bayesianas e algoritmo de *clustering*. As redes bayesianas criam um modelo a partir do conjunto de treinamento. Basicamente esse modelo tem em cada nó uma estrutura semelhante a uma árvore de decisão. As arestas entre os nós são informações dos consumidores (Bezerra, 2004). O algoritmo de *clustering*

(agrupamento) reúne usuários com perfis similares em grupos (clusters) bem definidos. Isso retorna ao usuário para que coexista em vários grupos simultaneamente com um fator de pertinência diferente para cada um deles (Bezerra, 2004). Santos (2008) utiliza o algoritmo de Similaridade de Cosseno como forma elucidar os problemas de similaridade entre usuários.

A Tabela 4 apresenta uma listagem que reúne a similaridade dos itens de interesse entre perfis. Com a utilização de um algoritmo que efetue o cálculo da similaridade entre Filtragem Colaborativa é possível mensurar quais perfis de usuário que apresentam hábitos em comum e quais aqueles diferenciados, de forma que os resultados levantados contribuam para uma recomendação correta e eficiente do produto.

Tabela 4: Recomendação baseada em Filtragem Colaborativa (Cazella, 2006).

Usuário	Produto 1	Produto 2	Produto 3	Produto 4
Paulo	X		X	X
Maria	X	X		
Renato		X	X	
Carlos				X

Os pesquisadores (Sarwar, 2000) e (Massa, 2004) listaram todas as principais deficiências no processo de Filtragem Colaborativa (algumas já citadas anteriormente) que ainda persistem:

- **O problema do primeiro avaliador**: não existe maneira de um determinado item ser recomendado para o usuário por filtragem colaborativa até que mais informações sobre o item sejam obtidas através de outro usuário. Sendo necessário realizar a avaliação inicial do item. A dificuldade está na obtenção desta primeira avaliação em qual estratégia a se adotar.

- **Similaridade**: Sendo assim, caso um usuário tenha hábitos que variem do padrão que o algoritmo de similaridade baseia-se, o Sistema de Recomendação terá dificuldades para encontrar outros usuários com “gostos” similares.

- **Novo Usuário**: Quando um usuário é cadastrado no sistema, ainda não existem avaliações realizadas por ele, portanto, seu perfil de avaliações está vazio. E para as recomendações serem realizadas é necessário encontrar os vizinhos mais próximos do usuário, e como ele ainda não realizou avaliações, não existirão vizinhos. A este problema é dado o nome de “Problema do Novo Usuário”. Também é referenciado na literatura como “*Cold Start User*”.

- **Novo item**: as recomendações são feitas baseada no perfil do usuário e de seus vizinhos mais próximos quanto a uma comunidade. Os itens de melhor avaliação dos vizinhos são recomendados ao usuário. Entretanto, quando um novo item aparece no sistema e não faz parte do perfil de nenhum usuário, não é possível realizar recomendações sobre ele.

- **Escalabilidade**: quando o volume de usuários, itens e avaliações são muito grandes, os sistemas que efetuam o algoritmo da vizinhança (“*k-nearest-neighbor*”) podem apresentar um tempo de resposta inaceitável.

- **Esparsialidade**: quando o número de itens na base de dados vai aumentando, diminuem as chances dos usuários possuírem itens similares e acarreta em uma vizinhança pobre para as recomendações.

- **Superespecialização**: esta limitação vem a tona quando o sistema só consegue recomendar itens similares àqueles que o usuário já avaliou, impossibilitando a atualização de novas recomendações.

5.2.3 Filtragem de Informação Híbrida

A abordagem de Filtragem de Informação Híbrida procura combinar as Filtragens Baseada em Conteúdo e Filtragem Colaborativa para construção de uma única técnica que busca a solução das limitações que ambas apresentam no processo de recomendação. A idéia é criar um sistema que possa melhor atender às necessidades do usuário (Herlocker, 2000), (Ansari, 2000).

A Figura 15 ilustra o potencial em reunir as duas técnicas de filtragem e os resultados que podem ser esperados.

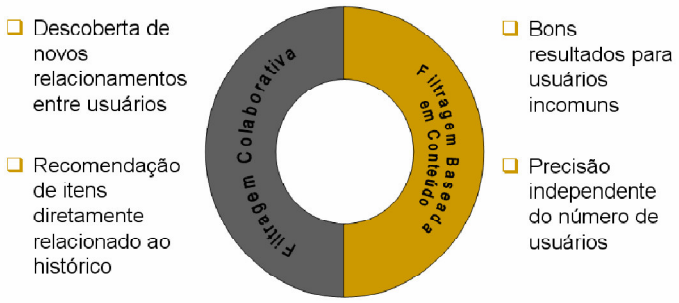

Figura 15: Representação Gráfica da Filtragem Híbrida e suas vantagens (Cazella, 2006).

O autor Montaner (Montaner et al., 2003), destaca que existe um quarto tipo de técnica de Filtragem de Informação denominada de Filtragem Demográfica. A filtragem demográfica utiliza a descrição de um indivíduo para aprender com o relacionamento entre um item em particular e o tipo de indivíduo que poderia vir a se interessar, gerando um tipo específico de perfil de usuário. Por conseguinte é criado um conjunto de perfis com aquele padrão e por fim caracterizá-los demograficamente.

Segundo Burke (2002), existe ainda as técnicas de Filtragem Baseada em Utilização e Filtragem Baseada em Conhecimento. Ambas não se preocupam com a avaliação do usuário, seus tópicos de interesse ou histórico

do perfil do usuário. A Filtragem Baseada em Utilização avalia a utilidade de cada objeto para o usuário, enquanto a Filtragem Baseada em Conhecimento busca as reais necessidades através de seus interesses e nas suas preferências para efetuar a recomendação. Para construir um sistema com estas duas técnicas, seria necessária uma representação mais detalhada e em longo prazo das necessidades do usuário para então suportar como um item em particular encontra uma necessidade particular.

O pesquisador Adomavicius (2005) apresenta na Tabela 5 uma proposta de classificação das abordagens de recomendação associados aos algoritmos de Mineração de Dados consolidadas neste âmbito, e divididos entre dois campos: Heurística e Modelo. Onde o autor define o campo da Heurística como as técnicas de responsáveis pelas predições de avaliações, baseando-se em toda a coleção de avaliações feitas pelos usuários aos itens; enquanto as técnicas baseadas em Modelo utilizam a coleção de avaliações para aprender o modelo, o qual será utilizado para realizar predições de avaliações. A classificação apresenta os algoritmos de Mineração de Dados mais utilizados.

Tabela 5: Abordagens de Recomendação x Mineração de Dados (Cazella, 2006).

Abordagens de Recomendação	Algoritmos de Mineração de Dados	
	Baseado em Heurística	**Baseado em Modelo**
Baseada em Conteúdo	TF-IDF (recuperação de informação); Agrupamento.	Classificadores Bayesianos; Agrupamento; Árvores de Decisão; Redes Neurais.
Colaborativa	Vizinhança mais próxima, Similaridade Cosseno, Correlação, Teoria dos Gráficos.	Redes Bayesianas, Agrupamento, Redes Neurais, Regressão Linear, Modelos Probabilísticos.
Híbrido	Combinando componentes baseados em conteúdo e colaborativos: Combinação linear de avaliação previstas; Esquemas variados de votação; Incorporando um componente como parte da heurística de outro.	Combinando componentes baseados em conteúdo e colaborativos: Incorporando um componente como parte de um modelo em outro; Construindo um modelo unificado

Este trabalho de Dissertação procura solucionar algumas das limitações citadas anteriormente nos tópicos de Filtragem Baseada em Conteúdo e Filtragem Baseada na Colaboração, através da criação de uma abordagem híbrida de recomendação que ao final gere uma Recomendação Social com possibilidade de encontrar dados de geolocalização agregado ao conteúdo do que será recomendado ao usuário; abordagem esta que será discutida no próximo capítulo.

6. METODOLOGIA PARA A RECOMENDAÇÃO SOCIAL

A aplicação do referido Trabalho de Dissertação tem como proposta a utilização de técnicas Mineração de Dados em Texto direcionados para a Recomendação Social em redes sociais que trabalhem com dados geolocalizados. Neste contexto, com a utilização das técnicas de Mineração de Dados, foi desenvolvida uma metodologia que busca diminuir as limitações da recomendação social e aplicar em uma Rede Social Baseada em Localização já consolidada.

6.1 A Recomendação Social

A recomendação vai desde interesses pessoais, produtos, serviços ou outros usuários (também chamada de recomendação social). A implicação social acompanha os sistemas de recomendação, pois além de agregar as recomendações por um conjunto de usuários ou serviços requisitados em comum similaridade; seria bem melhor formular agregações personalizadas em que a grande variedade de “gostos” diferenciados enriquecerá a recomendação.

Tendo em vista o papel do usuário em uma Rede Social como não somente um consumidor da informação, mas também manipulador dela; cabe ao sistema de recomendação ter como objetivo principal reduzir esta sobrecarga de informação através da seleção do conteúdo relevante em detrimento as preferências de cada usuário. Segundo Schafer (1999), os Sistemas de Recomendação são um complemento do processo social, no qual conta-se com conselhos ou sugestões de outras pessoas. Por exemplo, a recomendação social sugere aos usuários: *hobbys*, músicas, livros, lugares, notícias, ou ainda outras pessoas.

A recomendação pode acontecer em vários caminhos: contatos em redes sociais, músicas, filmes, livros, roupas, *tags*[2], restaurantes, *papers, e-learning,* roupas, hotéis e etc. Esquematicamente os Sistemas de Recomendação podem trabalhar desta forma:

- Prediz o quanto você pode gostar de certo produto ou serviço.
- Sugere uma lista de itens ordenada de acordo com seu interesse.
- Sugere uma lista de usuários ordenada para um produto ou serviço.
- Explica a você o porquê esses itens foram recomendados.
- Ajusta a predição e a recomendação baseado em seu *feedback* e de outros.

Segundo King (2010), a Recomendação social envolve a investigação de inteligência coletiva por meio de técnicas computacionais tais como: a Aprendizagem de Máquina, **Mineração de Dados**, Processamento de Linguagem Natural, etc. Onde o diferencial da Recomendação Social perante outras áreas de estudo é poder agregar dois aspectos: informação relevante e personalizada, e a informação relacional da Rede Social (entre os perfis de usuários. Estas duas fontes de informação social apresentam muitas aplicações que reúne teorias existentes, modelos, algoritmos e aplicações para o processamento da informação; tornando mais rica a recomendação social feita ao usuário. O trabalho de Lops (et al. 2011) apresenta uma proposta de Recomendação Social de trabalhos acadêmicos para os perfis cadastrados na rede *LinkedIn* com base no conteúdo existente em cada perfil e nas relações da Rede Social. A Figura 16 apresenta a arquitetura desta idéia que justifica a expansão dos relacionamentos provenientes dos interesses em comum que podem ser descobertos no ato da recomendação.

Enquanto a pesquisa conduzida por Argyriou (et al., 2011) apontam para o dinamismo da personalização de perfis como forma de garantir uma

[2] Palavra-chave que é relevante e associada a alguma informação. Sendo um recurso encontrado em muitos sites de conteúdo colaborativo e por essa razão, o termo "*tagging*" associa-se com a Web 2.0 (Fonte: http://pt.wikipedia.org/wiki/Tag_(metadata))

recomendação diversificada, conforme a Figura 17 destaca uma referência a todos os parâmetros (localização, itens de interesse, Rede Social, portabilidade de acesso e a rede de acesso) que são relacionados com a experiência do usuário personalizado. Em uma Recomendação Social há a possibilidade de uma crescente atualização dos dados que serão recomendados a partir do comportamento do usuário e do compartilhamento das informações.

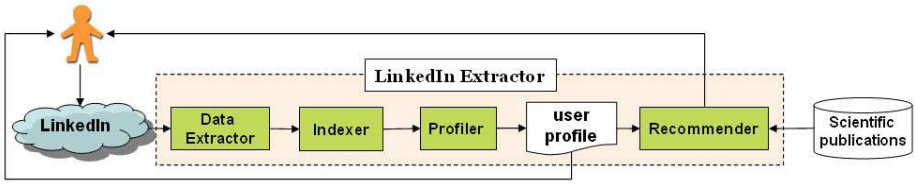

Figura 16: Arquitetura conceitual do *extractor* para *LinkedIn* (Lops, et al. 2011).

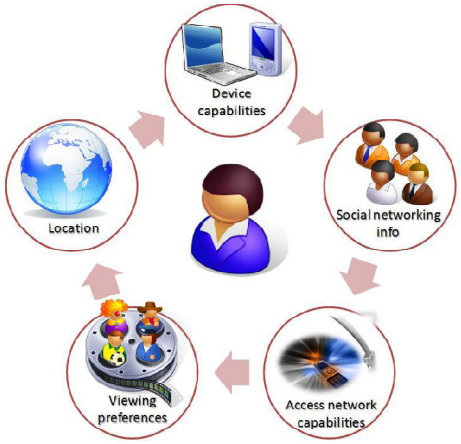

Figura 17: Personalização da experiência do usuário (Argyriou, et al. 2011).

O pesquisador King (2010) declara quais aspectos da Recomendação Social que ainda carecem de investigação:

1. A teoria de um melhor modelo formal e sobre as *cyber* interações sociais seria importante para que o futuro das interações sociais possa ser estimado.

2. Melhores algoritmos para mineração espacial existente (relacional) e temporal de dados com eficiência é necessário. Em particular, formas de lidar

com informações parciais e incompletas em sistemas: como sistemas de recomendação, *tagging systems,* uso de termos, etc; assegurando respostas mais precisas.

3. Análise de algoritmos do ponto de vista de escalabilidade. Como as Redes Sociais podem envolver com a complexidade do indivíduo e suas relações em comunidade, algoritmos devem calcular de forma eficiente e escalável.

4. Segurança e questões de privacidade são de grande preocupação na web, especialmente em Redes Sociais. Teorias e algoritmos para proteger informações pessoais são importantes quando as relações são facilmente criadas e dificilmente eliminadas.

5. Por último, uma questão interessante e altamente discutida é a monetização de interação social ou recomendação. Ou seja, encontrar maneiras para fazer ganhos financeiros com a Recomendação Social.

6.2 Contextualização e Trabalhos Relacionados

Inicialmente, para agregar com qualidade um estudo a respeito da Recomendação Social no que se propõe a pesquisa de Dissertação, foram analisados alguns trabalhos relacionados ao tema da pesquisa em dois âmbitos: o uso das técnicas de Mineração de Dados com propósito de Recomendação Social; e as pesquisas de tratamento dos dados direcionadas as Redes Sociais Baseadas em Localização. E a partir destes trabalhos analisados, enumerar os pontos fortes e desvantagens em cada pesquisa.

Lampe (2007) ressalta a importância de um perfil em uma Rede Social e como a se propaga as relações e conexões entre outros usuários. Foi elaborada uma coleta de dados visando a modelagem preditiva a cerca das articulações de amizade e um estudo a respeito das formações de comunidades on-line que agregam a interação de perfis com hábitos similares. No entanto, nenhuma ferramenta prática foi demonstrada.

Os pesquisadores Oka e Matsuo (2008) apresentaram a criação de uma Rede Social denominada *PolyPhonet* que utiliza técnicas de Mineração de Dados para extração de conhecimento acadêmico em bases de dados como *blogs* e *Wikipedia.* A pesquisa utilizou técnicas de *Web* Semântica através do algoritmo de similaridade Coeficiente de Sobreposição (uma extensão do Coeficiente de *Jaccard*) tendo como parâmetros da rede: perfis, suas relações e palavras-chaves. Sua pesquisa foca-se apenas nos relacionamentos e na abordagem colaborativa da co-ocorrência dos parâmetros, esquecendo o conteúdo dos perfis.

Nos trabalhos de Bhatia (2008) é proposto uso de um algoritmo de *Clustering* Largura-Primeiro para efetuar a Mineração de Dados utilizando uma abordagem estatística. O tempo de resposta do algoritmo torna-se uma limitação e sua metodologia de base estatística permite somente a extração de comunidades em uma Rede Social.

O pesquisador Motoyama (2009) concentrou seus estudos na construção de um sistema de busca eficaz para procura de outros usuários de das redes sociais: *Facebook* e *MySpace*. Entretanto, utiliza técnicas de teoria da aprendizagem e processamento de linguagem natural apenas para uma Filtragem Baseada em Conteúdo. Verificando o grau de eficiência na compatibilidade dos indivíduos em parâmetros biográficos.

Bartal (2009) apresenta o uso de técnica de Mineração de Dados em Texto junto a Análise das Redes Sociais como forma de efetuar previsões e comportamentos das interações de uma Rede Social utilizando Modelo Vetorial e Processamento de Linguagem Natural. No entanto, sua pesquisa não se aprofunda na riqueza de dados que podem ser minerados, se situando apenas na modelagem preditiva das mudanças em uma Rede Social.

Nas pesquisas de Romsaiyud (2011) foi proposto um Algoritmo de Lógica Fuzzy para efetuar a Mineração de Dados em uma Rede Social com o objetivo de extrair padrões comportamentais dos usuários. O trabalho somente centraliza a mineração de dados nas mensagens inseridas como forma de compreender os padrões psicológicos e sociológicos dos usuários.

A partir do que foi destacado anteriormente percebe-se que a utilização das técnicas de Mineração de Dados foi direcionada para modelagem preditiva e detecção de padrões do comportamento do usuário utilizando abordagens que ora prezavam pelo conteúdo dos perfis e itens da Rede Social e em outro momento procuravam descobrir os motivos das interações dos perfis.

Antes da análise dos trabalhos relacionados ao estudo das Redes Sociais Baseadas em Localização – LBSN's (*Location-Based Social Network*) elaborou-se um estudo a respeito das LBSN's mais utilizadas atualmente – sem a pretensão de esgotar todas LBSN's disponíveis para a pesquisa – tais como: *Google Places, Foursquare, MeetMoi, Google Latitude* e *Gowalla*; onde a Tabela 6 abaixo apresenta um resumo das descrições das principais LBSN's.

Segundo o pesquisador Jin (2012), a partir do crescimento das Redes Sociais Baseadas em Localização o foco das pesquisas de grande parte dos estudiosos neste ambiente tem sido a análise comportamental das atividades do usuário, seus relacionamentos com outros usuários e a sua mobilidade dentro da Rede Social buscando compreender a modelagem preditiva do perfil do usuário. Por exemplo, os usuários tendem a fornecer seus dados buscando muitas vezes demonstrar sua auto-representação, em detrimento a sua segurança ou privacidade (Cramer, 2011).

Segundo Vasconcelos (et al., 2012), os trabalhos com LBSN's analisados se dividem entre três campos: caracterização das LBSN's; propriedades sociais contrapondo as propriedades geográficas e análise de influências da Rede Social. O tópico de estudo referente a caracterização das Redes Sociais Baseadas em Localização apresentam alguns trabalhos que merecem destaque:

- Os autores Li and Chen (2009) utilizaram duas técnicas de agrupamento para modelagem do comportamento de usuários na Rede Social *Brightkite*. Uma das abordagens classificou os usuários quanto à mobilidade dos padrões de atualização (*checkins*, fotos e comentários) e a outra técnica agrupou esses mesmos usuários levando em conta também os aspectos sociais.

- A pesquisa de Noulas et al. 2011 procurou resultados mais sólidos, baseando na criação de apenas um algoritmo de agrupamento espectral para reunir os usuários de uma Rede Social através dos padrões de *checkins* feitos. O objetivo era dividir os usuários em grupos a fim de identificar comunidades e caracterizar o tipo de atividade em cada região de uma cidade.
- O pesquisador Gambs (2011) desenvolveu um trabalho de análise comparativa dos critérios de privacidade na LBSN's do compartilhamento dos dados de geolocalização.
- O trabalho de Ferrari (et al., 2011) procura identificar padrões de mobilidade urbana associada a busca de similaridade social entre os usuários. Onde foi adotado Modelo Probabilístico na extração dos dados e clusterização para agrupar as pessoas.

Quanto a análise entre propriedades sociais e propriedades geográficas das LBSN's outras pesquisas a serem listadas são:

- O pesquisador Mao Ye (2011) fez uma abordagem voltada para um levantamento das características dos usuários e características das regiões propondo uma técnica de filtragem colaborativa para os "*friends*" da Rede Social *Foursquare.* Procurando efetuar uma análise dos laços sociais junto as regiões favoritas de cada usuário.
- O pesquisador Cho (et al., 2011) trabalhou com modelagem dos padrões de mobilidade nas redes do *Gowalla*, *Brightkite* e *traces* de telefones celulares.
- Enquanto Cheng (et al., 2011) fez uma análise quantitativa sobre os padrões de mobilidade humana levando em consideração propriedades espaciais, temporais, sociais e textuais dos *checkins* em diversas redes LBSNs.
- O pesquisador Jie Bao (2012) desenvolveu uma recomendação social na LBSN *Foursquare* onde procurou compreender o comportamento dos usuários baseadas nas preferências pelos locais por eles visitados através de um algoritmo de seleção do melhor candidato (usuário) como forma de elaborar uma recomendação personalizada.

Por fim, o terceiro aspecto a ser tratado é a análise de influências nas LBSN's, onde poucos trabalhos da área abaixo são listados:

- O pesquisador Carlone (2011) desenvolveu um sistema que utiliza uma abordagem de Mineração de Dados Textual para lingüística utilizando técnicas de processamento de linguagem natural como forma de extrair palavras individuais em opiniões destacadas em uma LBSN.

- Melià-Seguí (2012) abordou o estudo do comportamento do usuário e sua atividade na Rede Social do *Foursquare,* desde os aspectos quantitativos (número de usuários atuantes) ao número de *checkins* realizados. A idéia era melhorar o tempo de recomendação para os usuários.

- Os trabalhos referentes a Vasconcelos (et al., 2012) focaram apenas nas análises dos padrões de comportamento e na interação dos usuários no *Foursquare* a partir dos comentários feitos pelos os usuários; agrupando estes em apenas dois grupos específicos.
- Costa (2013) apresenta pesquisa voltada para a detecção de mensagens *spams* na LBSN brasileira Apontador através de técnica de classificação de Mineração de Dados: algoritmo *Random Forest.* O objetivo é a prevenção de ataques de *spammers* (mensagens eletrônicas não-solicitadas enviadas em massa).

Os trabalhos com as Redes Sociais Baseada em Localização ainda são escassos, sobretudo no quesito a análise de influências. As pesquisas relacionadas anteriormente são recentes e ainda carecem de estudos voltados a filtragem do conteúdo dos dados (estruturados e desestruturados) que são compartilhados nas LBSN's de forma que seja melhorado o desempenho do que é extraído e posteriormente recomendado socialmente aos usuários.

Tabela 6: Quadro-resumo das especificações das principais LBSN's.

LBSN	*Google Places*	*Foursquare*	*MeetMoi*	*Google Latitude*	*Gowalla*	*Brightkite*
Propósito	Cadastro de propriedades e empresas e compartilhando a localização do perfil na web de forma gratuita.	Compartilha a localização dos usários, lista e Recomenda locais favoritos para visitar a partir dos comentários dos usuários sobre os locais que visitou.	Recomenda perfis de pessoas para um relacionamento baseado no grau de proximidade e similaridade de perfis em um local (*Location-Based Modile Dating*).	Compartilha a localização dos usuários, além de permitir comentários.	Compartilha a localização dos usários, fotos, registros de viagens e Recomenda para os usuários sobre os locais que visitou.	LBSN comercial que permite que os usuários compartilhem sua localização, *posts*, e façam o upload de fotos com as configurações de privacidade ajustáveis.
Pontos Fortes	Permite ser integrado a outra Rede Social: *Google+*. Cria um ranking de dados a ser utilizado em buscas relativas a empresas e usuários-clientes.	Facilidade de Uso e Permite ser integrado a outras Redes Sociais: *Facebook e Twitter.* Alta difusão entre os dispositivos móveis e apresenta também um mecanismo de busca simples.	Facilidade de Uso. Interoperabilidade de operadoras de rede. Apresenta um bom desempenho através dos dispositivos móveis. Quebra com a estrutura de comunidades fechadas em uma Rede Social.	Facilidade de Uso. Alta disponibilidade em dispositivos móveis.	Facilidade de Uso e Permite ser integrado a outras Redes Sociais: *Facebook e Twitter.* Alta difusão entre os dispositivos móveis e apresenta também um mecanismo de busca simples.	Permite uma configuração da privacidade pelo usuário quanto as localizações. Um usuário pode descobrir nas proximidades pessoas e procurar os seus fluxos de atividades públicas.
Pontos Fracos	Usa apenas Abordagem Baseada em Conteúdo na Recomendação Social.	Usa apenas Abordagem Colaborativa para a Recomendação Social. A Recomendação somente é eficiente quanto maior o território do local.	Usa apenas Abordagem Baseada em Conteúdo para recomendação social. É alto dependente dos dispositivos móveis.	Prevalece apenas Abordagem Colaborativa na Recomendação Social. Anomimização dos dados é ineficaz.	Usa Abordagem Baseada em Conteúdo para a Recomendação Social. A Recomendação é eficiente quanto maior o território do local.	O uso comercial da LBSN atrapalha sua difusão. Prevalece apenas Abordagem Colaborativa na Recomendação Social.

Conforme destacado no Capítulo 4 deste Trabalho de Dissertação, a Rede Social do *Foursquare* foi escolhida para atuar como a área de domínio na criação e implementação da metodologia de Mineração de Dados para a Recomendação Social em Redes Sociais Baseadas em Localização tendo a área de análise de Influências como o campo da pesquisa com LBSN.

A justificativa se deve aos respectivos aspectos: a grande utilização desta LBSN em todo mundo por diversas pessoas em várias localidades, além de sua gratuidade; utiliza apenas a abordagem colaborativa como técnica de filtragem de informações carecendo de técnicas de filtragem baseada em conteúdo, ou ainda uma abordagem híbrida que possa trazer melhores resultados na recomendação social; quanto menor a localidade mais complexa torna-se a recomendação social feita aos usuários; a recomendação da Rede Social *Foursquare* apresenta as limitações típicas dos Sistemas de Recomendação como o *Cold Start Problem*, problemas de Similaridade (Ovelha Negra) e Superespecialização.

6.3 Construção da metodologia Híbrida de Mineração de Dados

A construção da metodologia de Recomendação Social para as Redes Sociais Baseadas em Localização procura atingir uma abordagem de recomendação Híbrida utilizando tanto a técnica de Filtragem de Informação Baseada em Conteúdo quanto à técnica de Filtragem de Informação Baseada na Colaboração.

A partir das pesquisas das técnicas de Mineração de Dados junto aos trabalhos relacionados com tema foram escolhidos dois métodos de Mineração de Dados visando filtragem de informações textuais: o Modelo Vetorial com o algoritmo TF-IDF (*Term Frequency Inverse Document Frequency*) e Coeficiente de Agrupamento (Método de Clusterização) Similaridade de Cosseno.

6.3.1 Abordagem utilizando algoritmo TF-IDF (*Term Frequency Inverse Document Frequency*)

Em uma Rede Social Baseada em Localização é comum encontrarmos informações textuais, a serem extraídos, presentes nas descrições dos perfis, descrição de itens, comentários dos usuários, listas de locais visitados ou qualquer outro recurso que apresenta informação textual que possa ser compartilhado na LBSN. Estas informações geralmente se apresentam incorporadas através de dados estruturados ou desestruturados, onde um coletor de dados tratará de reter os corpus (coleção) de documentos que são os responsáveis pela contribuição da recomendação social.

Tendo em vista o que foi comentado no Capítulo 3 da presente Dissertação, a proposta do algoritmo TF-IDF (Salton, 1988) é poder trabalhar com toda esta informação textual coletada (sabendo que os corpus são visto como vetores de palavras em um espaço vetorial) da LBSN e gerar um expressivo *ranking de* pontuações normalizadas dos termos presentes nos documentos. O cálculo da Freqüência dos Termos multiplicado a Freqüência Inversa dos Termos em um documento da LBSN contribui na medição da similaridade através de uma pontuação que contabiliza ambos os fatores de freqüência, pois os melhores termos de indexação (os que apresentarão maior peso) são aqueles que ocorrem com uma grande freqüência em poucos documentos. Esta estimativa contribui na relevância do que deve ser adotado para recomendação social.

Por exemplo, para entendimento deste raciocínio os seguintes corpus de três documentos abaixo poderiam estar presentes em uma LBSN:

a. “O restaurante Panela de Mina é o roteiro ideal para a Familia Maranhense. Apresentando um cardápio variado para que todos da Familia Maranhense saboreiem”

b. “A Familia Brasileira tem ido mais vezes ao cinema”

c. “O shopping do Rio Anil tem apresentado inúmeras possibilidades de diversão para a familia”

É normal que a freqüência de um termo possa ser simplesmente representada como um número de vezes em que ele ocorre no texto, entretanto é bem mais comum que seja normalizado considerando um número total de termos no texto. Isto contribui para que a pontuação geral considere também a extensão do documento relativa a freqüência de um termo.

Ou seja, o termo “família” presente nos três documentos acima (considerando depois de normalizado apenas por letras minúsculas) ocorre duas vezes no documento *a* e apenas uma vez no documento *b.* Assim imagina-se que o corpus *a* produziria uma pontuação bem mais alta que o corpus *b,* no entanto, este não é o único critério para medição da freqüência. Normalizando o comprimento dos documentos, o corpus *b* terá uma freqüência maior para o termo “família” que o corpus *a.* Onde o comprimento de cada um respectivamente seria: *a* – 2/24; *b* – 1/9; ainda que ocorra mais vezes no corpus *a* (duas vezes em vinte dois termos) o corpus *b* é mais curto (uma para nove termos).

Neste sentido, um recurso ideal para a pontuação de uma consulta com mais de dois termos como, por exemplo, “família maranhense” é somar a pontuação da freqüência dos termos para cada um dos termos de consulta em cada documento, e retornar os documentos classificados pela pontuação somada pela freqüência dos termos.

Utilizando ainda como exemplo os três documentos apresentados anteriormente é listado as seguintes pontuações acumuladas na Tabela 7 abaixo onde o corpus *a* é identificado como o resultado esperado.

Tabela 7: Pontuações corpus x freqüência do termos.

Documento	TF(família)	TF(maranhense)	Soma
Corpus *a*	2/24	2/24	4/24 (0,1666)
Corpus *b*	1/9	0	1/9 (0,1111)
Corpus *c*	1/14	0	1/14 (0,0714)

Apesar de o resultado ser eficiente ainda ocorre alguns problemas na pontuação acumulada, pois ela analisa a freqüência dos termos de cada documento como uma coleção não-ordenada de palavras. Consultas como, por exemplo, “Maranhense Família”, “Sr. Família Maranhense” ou similar, retornariam o mesmo valor de pontuação acumulada para o termo “família maranhense”, ainda que nenhuma destas construções de termos (*token*) estivesse presente nos documentos. Ou seja, a pontuação gramatical que acompanha os textos junto ao contexto que acompanha os *tokens* de interesse não é levada em consideração.

Esta complexidade considera que somente a freqüência dos termos não ajuda quando ocorrem palavras muito comuns em documentos. Por exemplo, o termo “para Família Maranhense” contém a *stopword* “para” que distorce a pontuação geral da freqüência dos termos em favor do corpus *a,* pois o termo “para” apresenta-se duas vezes no corpus *a* além dos outros termos que se busca em detrimento aos outros corpus onde o termo “para” se apresenta apenas no corpus *c.*Mesmo que a contextualização das frases diga o contrário mas a pontuação acumulativa não prever este problema.

O cálculo da métrica da freqüência inversa dos termos para um documento fornece os recursos necessários para normalização de um corpus onde considera a presença de termos usuais de uma coleção de documentos verificando o número total de documentos que um termo de consulta está presente. O objetivo é que esta métrica produza um valor mais alto caso um termo seja relativamente incomum o que ajuda a lidar com o problema das *stopwords*[3]*.* Por exemplo, uma consulta por “família” no corpus dos três documentos de exemplo retorna uma pontuação de freqüência inversa do documento mais baixa se for comparar com uma consulta por “cardápio” devido a o primeiro termo da consulta estar presente em todos os documentos, enquanto “cardápio” em apenas um. Com os estudos de freqüência os seguintes critérios são listados: a Freqüência dos Termos presentes em um documento, o comprimento do documento e a unicidade geral dos termos nos

[3] Também conhecido como palavras de ligação ou de parada. Palavras pobres para discriminar ou identificar o conteúdo de um documento e que são muitos freqüentes não coleção do documento.

documentos que são coletados. Ou seja, a união destes critérios denota o algoritmo TF-IDF. Para exemplificar toda esta análise do algoritmo TF-IDF, será ilustrado o cálculo em um hipotético documento (não longo mais suficiente para o teste) de comentários de uma LBSN que apresenta exemplos de corpus (alguns já citados anteriormente) e a forma que este pode gerar resultados consideráveis.

a) "O restaurante Panela de Mina é o roteiro ideal para a Familia Maranhense. Apresentando um cardápio variado para que todos da Familia Maranhense saboreiem"

b) "A Familia Brasileira tem ido mais vezes ao cinema"

c) "O shopping do Rio Anil tem apresentado inúmeras possibilidades de diversão para a familia"

d) "As lojas Americanas tem o presente para a melhor familia"

e) "Não acho os terminais da integração ideais para o tráfego de uma familia"

Considerando o documento a idéia é a busca por três termos distintos: "familia", "melhor familia" , "a familia maranhense". A Tabela 8 apresenta as pontuações dos cálculos individuais em cima de cada termo de consulta exibindo a Freqüência dos Termos, a Freqüência Inversa dos Termos e o cálculo TF-IDF; quanto a relação dos cinco corpus do documento. Posteriormente, a Tabela 9 detalha uma outra possibilidade de resultados com soma das pontuações TF-IDF para cada corpus do documento. Por exemplo, para a consulta "melhor familia" e a "a familia maranhense" no corpus *b* os resultados são: 0 + 0.1111 = 0.1111; - 0.2899 + 0.1111 + 0 = 0.4009

A idéia que está inserida no cálculo do algoritmo TF-IDF é que com a multiplicação dos pesos de dois termos é possível a capacidade de produzir pontuações TF-IDF maiores para consultas mais relevantes ao que o usuário necessita do que para consultas menos relevantes. Entretanto alguns *stopwords* podem ser descartados em coleções de documento mais robusto (como o caso do termo "a" do exemplo), pois neste caso não apresentam uma forte modificação nos pesos.

Tabela 8: Quando de pontuações do cálculo TF-IDF.

Documento	TF (familia)	TF (maranhense)	TF (a)	TF (melhor)
Corpus *a*	0.0833	0.0416	0.1087	0
Corpus *b*	0.1111	0	0.2899	0
Corpus *c*	0.0714	0	0.1863	0
Corpus *d*	0.1	0	0.2609	0
Corpus *e*	0.0769	0	0	0.1
	IDF (familia)	**IDF (maranhense)**	**IDF (a)**	**IDF (melhor)**
Corpus *a*	1.0	2.6094	1.2231	2.6094
Corpus *b*	1.0	2.6094	1.2231	2.6094
Corpus *c*	1.0	2.6094	1.2231	2.6094
Corpus *d*	1.0	2.6094	1.2231	2.6094
Corpus *e*	1.0	2.6094	1.2231	2.6094
	TF-IDF(familia)	**TF-IDF (maranhense)**	**TF-IDF (a)**	**TF-IDF (melhor)**
Corpus *a*	0.0833	0.0416	0.1087	0
Corpus *b*	0.1111	0	0.2899	0
Corpus *c*	0.0714	0	0.1863	0
Corpus *d*	0.1	0	0.2609	0
Corpus *e*	0.0769	0	0	0.1

Tabela 9: Valores TF-IDF somados para a consulta.

Consulta	*a*	*b*	*c*	*d*	*e*
familia	0.0833	0.1111	0.0714	0.1	0.0769
melhor familia	0.0833	0.1111	0.0714	0.1	0.1769
a família maranhense	0.2336	0.4009	0.2577	0.3609	0.0769

6.3.2 Abordagem utilizando algoritmo Similaridade de Cosseno

O algoritmo TF-IDF fornece os recursos necessários para analisar os termos de consulta em um documento e assim quantificar os resultados mais próximos daquilo que o usuário busca. Entretanto algumas limitações quanto a similaridade dos documentos ainda são encontradas. Algoritmos de similaridade contribuem nesta proposta de buscar o documento mais semelhante ao que o usuário procura. Nesta pesquisa foi adotado o algoritmo de Similaridade de Cosseno (Adomavicius, 2005) (Santos, 2008) para o agrupamento dos documentos mais similares. Para entendimento do algoritmo de Similaridade Cosseno é necessário inicialmente um entendimento do Modelo de Espaço Vetorial neste contexto.

Conforme já destacado anteriormente pelo Modelo de Espaço Vetorial: o cenário é de um espaço multidimensional que contém um vetor para cada documento, e a distância entre dois vetores indica a similaridade entre os documentos. Ou seja, é possível representar uma consulta como um vetor, e encontrar os documentos mais relevantes através do seguinte comparativo: **vetores de documentos com a menor distância para o vetor da consulta**.

No contexto da Mineração de Dados Textuais o “vetor” é entendido como uma lista de números que expressa tanto a direção relativa a uma origem; quanto uma magnitude expressa a distância dessa origem. Portanto, um vetor pode ser ilustrado como um segmento de linha entre a origem e um ponto em um espaço N-dimensional desenhado como uma linha entre a origem e o ponto.

Por exemplo, um documento que se restringe a apenas dois termos: (“Carro, “Marca”) com um vetor correspondente (0.32, 0.58), onde os valores no vetor são atribuídos, tal quais as pontuações TF-IDF para os termos. Vislumbrando em um espaço vetorial, esse documento poderia ser representado em duas dimensões por um segmento de linha que se estenderia desde a origem (0,0) até o ponto (0.32, 0.58); como uma analogia ao plano X/Y

em que o eixo X seria Carro, o eixo Y seria Marca e o vetor de (0,0) até (0.32, 0.58) seria o documento em questão.

Visualizando um documento maior com um vetor de termos (“Carro”, “Marca”, “Ano”) em um espaço tridimensional, o cosseno do ângulo entre dois vetores é a métrica ideal para comparação dos vetores, recebendo também o nome de Similaridade de Cosseno.

A Figura 18 abaixo ilustra um documento visto em um espaço tridimensional.

	t_1	t_2	t_3
DOC_2	0.5	0.4	0.3

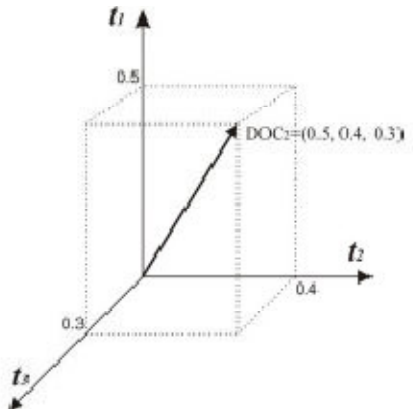

Figura 18: Representação do documento com seus pesos em espaço 3D.

Esta escolha pode justificada pela seguinte afirmação: quanto mais próximos estiverem dois vetores, conseqüentemente menor será o ângulo entre eles e desta forma maior o cosseno deste ângulo. Ou seja, dois vetores semelhantes teriam um ângulo de 0 grau e uma medida de similaridade de 1.0, enquanto dois vetores octogonais teriam um ângulo de 90 graus e uma medida de similaridade de 0.0. A Tabela 10 abaixo descreve este raciocínio.

Apesar de um exemplo simples de documento, o mesmo raciocínio cabe em documentos que contenham centenas ou mais termos em um espaço de três, dez ou mais dimensões, necessitando de uma análise de similaridade eficiente.

Tabela 10: Representação do Algoritmo Similaridade de Cosseno.

Formulação Matemática	Categorias para cada Formulação
$\overrightarrow{doc1} \bullet \overrightarrow{doc2} = \| doc1 \| \bullet \| doc2 \| \bullet \cos 0$	Dado (trigonometria)
$\frac{\overrightarrow{doc1} \bullet \overrightarrow{doc2}}{\| doc1 \| \bullet \| doc2 \|} = \cos 0$	Pela divisão
$\wedge \quad \wedge$ $doc1 \bullet doc2 = \cos 0$ $\wedge \quad \wedge$ $doc1 \bullet doc2 = \mathrm{Sim}\,(doc1, doc2)$ $\cos 0 = \mathrm{Sim}\,(doc1, doc2)$	Vetor Unitário Algoritmo de Similaridade Cosseno

Portanto, para utilizar o algoritmo Similaridade de Cosseno é necessário apenas produzir um vetor de termos para cada documento e computar o produto escalar dos vetores unitários para esses documentos; por isso os pesos proporcionados pelo algoritmo TF-IDF são reutilizados para este raciocínio. Por exemplo, usando como referência a consulta de termos e os pesos gerados pelo cálculo do TF-IDF na coleção de documentos da seção anterior elabora-se o cálculo de Similaridade de Cosseno, tais como:

- **Corpus *a*** apresenta um vetor de documento para o termo de consulta “familia”, “melhor familia” e “a familia maranhense” = (0.0833, 0.0833, 0.2336)
- **Corpus *b*** apresenta um vetor de documento para o termo de consulta “familia”, “melhor familia” e “a familia maranhense” = (0.1111, 0.1111, 0.4009)
- **Corpus *c*** apresenta um vetor de documento para o termo de consulta “familia”, “melhor familia” e “a familia maranhense” = (0.0714, 0.0714, 0.2577)
- **Corpus *d*** apresenta um vetor de documento para o termo de consulta “familia”, “melhor familia” e “a familia maranhense” = (0.10, 0.10, 0.3609)

- **Corpus *e*** apresenta um vetor de documento para o termo de consulta “familia”, “melhor familia” e “a familia maranhense” = (0.0769, 0.1769, 0.0769)

Conforme a representação matemática apresentada na Tabela 10 e após a pesquisa do usuário são gerados vetores de consulta em cima de cada documento (vetores de documento). O algoritmo de Similaridade de Cosseno procura efetuar a comparação de cada vetor documento a outro vetor documento. Os resultados obtidos na comparação estão descritos abaixo:

- Vetor 0 x Vetor 1 = 0.0043737287738697672,
- Vetor 0 x Vetor 2 = 0.0043806160919012127,
- Vetor 0 x Vetor 3 = 0.0043784718045359883,
- Vetor 0 x Vetor 4 = 0.28019216352279019,
- Vetor 1 x Vetor 2 = 2.7151774162348374e-09,
- Vetor 1 x Vetor 3 = 1.2880001509785188e-09,
- Vetor 1 x Vetor 4 = 0.33985500588848128,
- Vetor 2 x Vetor 3 = 2.6304447509062356e-10,
- Vetor 2 x Vetor 4 = 0.33990434514803569,
- Vetor 3 x Vetor 4 = 0.33988898770356968

A consulta de um espaço vetorial aponta exatamente a mesma operação que se usa para comparar a similaridade entre documentos; entretanto, com as pontuações geradas do algoritmo TF-IDF acaba se comparando o vetor de consulta e os vetores de documento. Os maiores resultados das comparações atestam a maior similaridade entre os documentos o que contribui para o agrupamento destes e uma melhor recomendação social. Por exemplo, o vetor de documento 0 e vetor de documento 3 apresentam um grau de similaridade fraca, enquanto os vetores de documento 1 e 4 apresentam um grau de similaridade bem maior. A Figura 19 ilustra a abordagem Híbrida de Mineração de Dados em Texto.

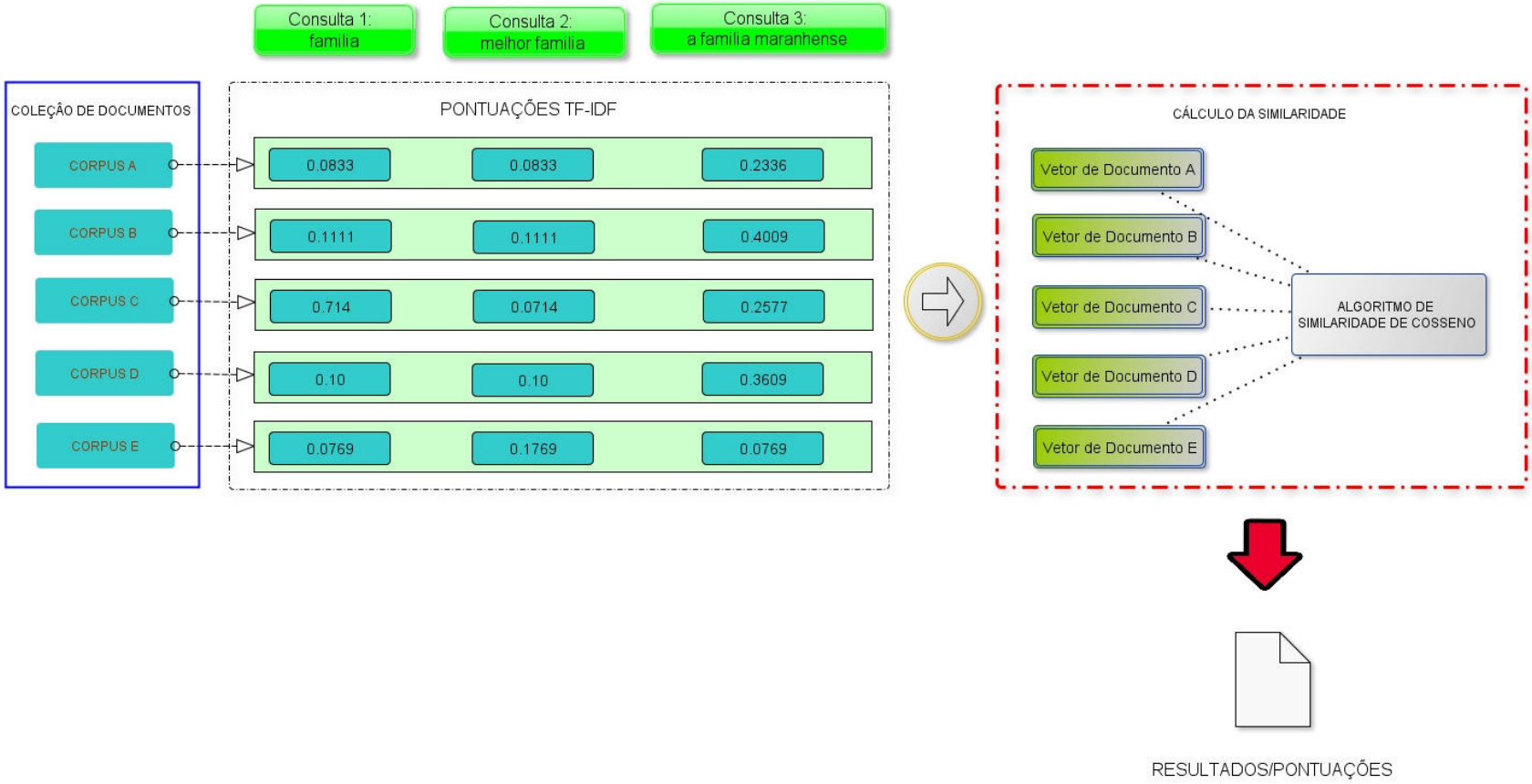

Figura 19: Representação Gráfica da Abordagem Híbrida de Mineração de Dados para a Recomendação Social.

As medidas de similaridade são medidas de distância e geralmente são usadas como pesos para encontrar os vizinhos mais próximos (Santos, 2008), ou seja, quão similares são as consulta dos usuários quanto aos resultados obtidos. Tendo em vista que o TF-IDF apresenta suas limitações a métrica da Similaridade de Cosseno procura solucionar algumas deficiências do cálculo da Freqüência dos Termos e Freqüência Inversa dos Termos propondo um cálculo mais refinado e com um agrupamento dos corpus com maior semelhança com o que o usuário busca nos documentos de uma LBSN.

6.4 Recomendação por *Tags*

A análise de Recomendação por *Tags* pode ser trabalhada por métodos de Filtragem Baseado em Conteúdo e métodos de Filtragem Baseados na Colaboração. Ou seja, as *tags* podem ser definidas para o item que se deseja encontrar, ou usadas anteriormente pelo usuário, ou ainda co-ocorrerem com as *tags* já descritas no ato da busca. Resumindo no seguinte cenário: uma recomendação de *tags* enviadas por usuários similares para itens similares (Camel, 2010). O pesquisador Heymann (2008) estudou a previsão de uma *tag* social e descobriu que as regras baseadas em *tag* de associação podem produzir muitos resultados de alta precisão. Enquanto Schenkel (2008) utilizou as *tags* como expansões semânticas para ajudar na busca social.

Neste Trabalho de Dissertação é proposto uma abordagem baseada em conteúdo para a filtragem dos dados na LBSN *Foursquare*, e depois através das técnicas Mineração de Dados Textuais fazer a Recomendação Social com extensão por *Tags*. A partir da definição de *tags* pelo usuário na rede social, são listados possíveis termos “candidatos” ao que o usuário busca. Com as técnicas de Mineração de Dados Textuais será feito a agregação e *ranking*, com objetivo de filtrar e recomendar socialmente os resultados mais similares ao usuário. O diferencial nesta pesquisa é a possibilidade de aumentar a capacidade da recomendação social através da associação de objetos de consulta entre informações textuais e semânticas. A Figura 20 apresenta um exemplo de raciocínio similar, nos trabalhos de Sigurbjörnsson (2008).

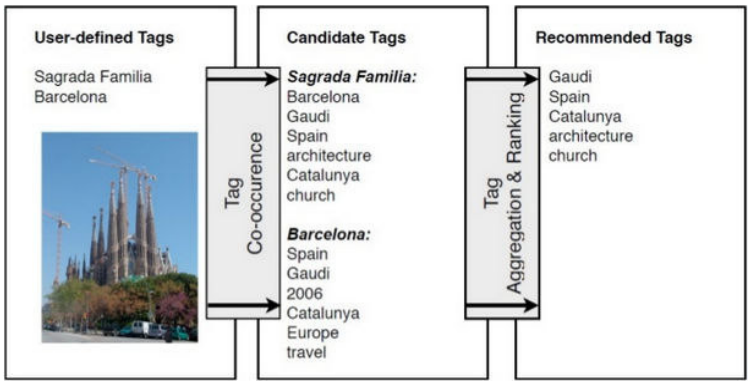

Figura 20: Representação da Recomendação por *Tags* (Sigurbjörnsson, 2008).

A Co-ocorrência de *Tags* torna-se um método ideal para uma recomendação social bem mais enriquecida ao que usuário busca de forma direta ou indiretamente. Pois, alguns recursos de uma Rede Social Baseada em Localização podem ser **considerados como *tags* ou conter *tags***, tais como: comentários dos usuários, listas de lugares, descrições em perfis, nomes de locais, entre outros. Outros pesquisadores adotaram metodologias de Recomendação por *Tags* que também contribuíram a este trabalho de Dissertação como referencial bibliográfico como: Zhao (et al., 2008) propôs calcular a similaridade de dois usuários com base na distância semântica em que suas *tags* definem itens comuns que marcou. Tso-Sutter (et al., 2008) estendeu os vetores de itens para os perfis de usuário e vetores do usuário para os perfis de itens com *tags* e depois construiu medições de vizinhança usuário / item com base nos perfis. Portanto, a área de Sistemas de Recomendação como *Tagging Social* (Folksonomia[4]) tornou-se ativa e crescente o tópico de estudos divididos em três campos: sugestões de *tags*, pesquisas sociais, e recomendações sociais (Kim, 2012). A Figura 21 abaixo apresenta o raciocínio da Recomendação por meio de *Tags* como forma de efetuar após a utilização do algoritmo Híbrido de Mineração de Dados o raciocínio de recomendação social.

[4] A Folksonomia é uma maneira de indexar informações. Enquanto na taxonomia clássica, primeiro são definidas as categorias do índice para depois encaixar as informações em uma delas (e em apenas uma), a Folksonomia permite a cada usuário da informação classificar com uma ou mais palavras-chaves, conhecidas como tags (Fonte: http://pt.wikipedia.org/wiki/Folksonomia)

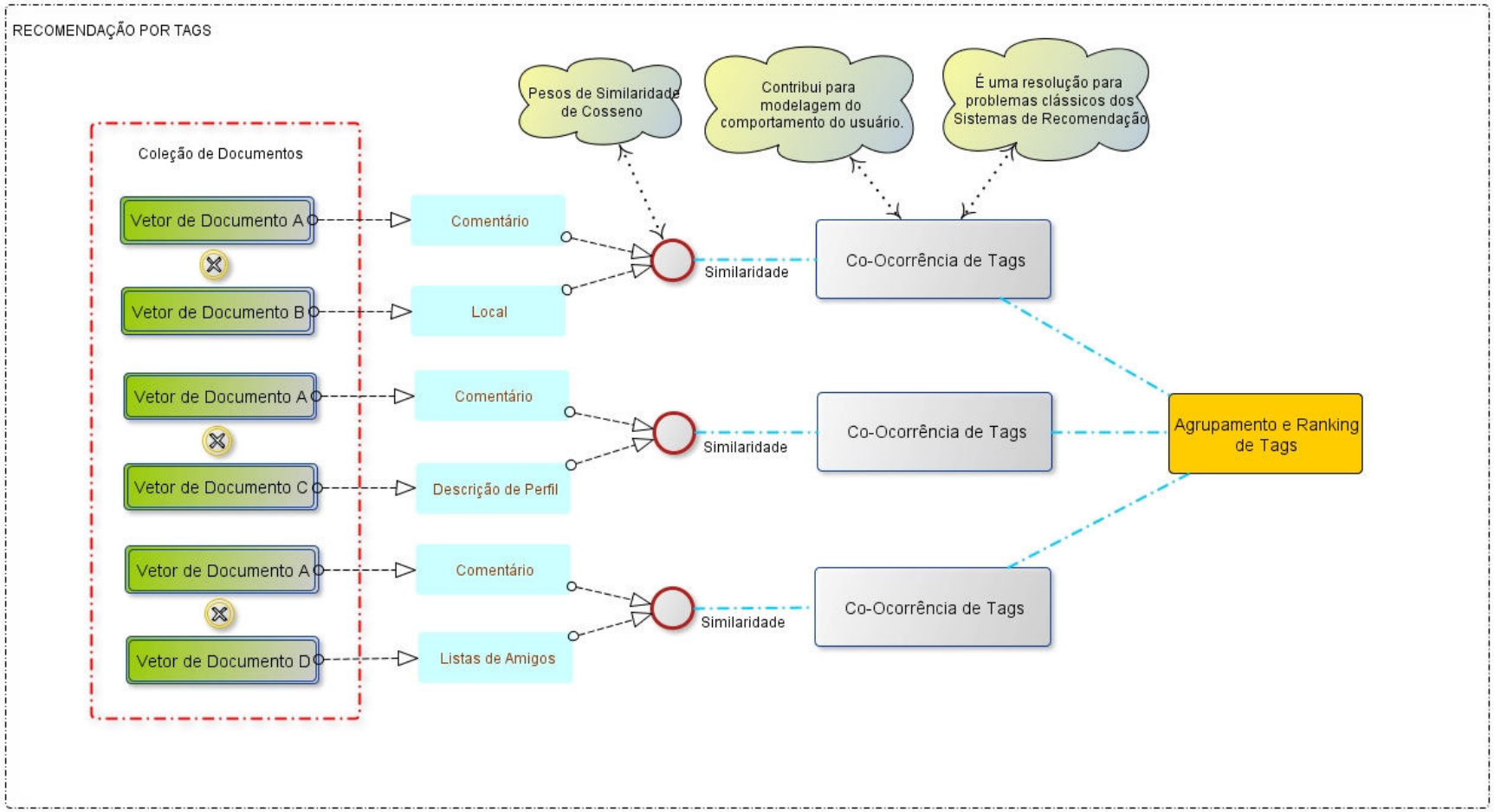

Figura 21: Proposta da Recomendação por *Tags*.

Entretanto após realizar as medições de similaridade entre os documentos é necessário agrupá-los em vizinhanças. Isso pode ser feito através de técnicas que trabalhem com o conceito de **Número de Vizinhos**. Onde se considera um critério de comparação de valores chamado Limiar de Similaridade (Burke, 2000). Ou seja, os documentos que tiverem um valor superior ao limiar serão definidos como vizinhos do usuário alvo. A desvantagem nesta técnica é a possibilidade de criar uma baixa vizinhança ou nenhuma vizinhança. Entretanto torna-se ideal a ser aplicado neste trabalho de Dissertação pela grande coleção de documentos extraído da LBSN.

Cada corpus do documento é contido em entidades que o caracterizam como dados desestruturados, ou seja, em formato de *tags.* As *Tags* carregam consigo o potencial de se relacionarem com outras *tags* fora do domínio da consulta agregando assim conhecimento (devido à co-ocorrência com outras *tags*) e trabalharem com os dados geolocalizados.

Com intuito de promover o agrupamento e ranking das *tags* encontradas no próprio corpus, de documento de acordo com metodologia descrita neste trabalho utilizam-se técnicas de Clusterização tal como: o algoritmo k-NN (*k-Nearest Neighbour*) também chamado de algoritmo do **Vizinho mais Próximo**. O uso da técnica k-NN no Agrupamento e *Ranking* das *Tags* busca a organização dos documentos em torno do vetor espaço das *Tags.* Este raciocínio é chamado de Similaridade de Contexto de *Tag* (Cattuto et al., 2008) onde um vetor de *Tags* é construído com base na análise comparativa das pontuações (pesos) dos vetores de documento. A Tabela 11 ilustra esta metodologia que tem como objetivo detectar as ligações semânticas entre as *tags* e obter uma melhor recomendação social. Onde o peso é da co-ocorrência entre *tags;* o "VetorTag" é construído contando quantas vezes a *Tag* co-ocorre com *Tag'* (Cattuto et al., 2008).

A colaboração das *Tags* contribui para identificação de padrões e as localizações das pessoas (inclusive as que têm algo em comum e estão próximas) que se relacionam sugerindo uma recomendação social padronizada. A Figura 22 ilustra esta técnica através da adaptação ao Grafo LBSN de Mao Ye (2011).

Tabela 11: Ilustração da Similaridade de Contexto de Tag.

	Tag2	Tag3	Tag4	Tag5	Tag6	Tag7	
Tag1	Peso	Peso	Peso	Peso	Peso	Peso	**VetorTag**
	Tag1	**Tag3**	**Tag4**	**Tag5**	**Tag6**	**Tag7**	
Tag2	Peso	Peso	Peso	Peso	Peso	Peso	**VetorTag**
	Tag 1	**Tag2**	**Tag4**	**Tag5**	**Tag6**	**Tag7**	
Tag3	Peso	Peso	Peso	Peso	Peso	Peso	**VetorTag**
	Tag1	**Tag2**	**Tag3**	**Tag5**	**Tag6**	**Tag7**	
Tag4	Peso	Peso	Peso	Peso	Peso	Peso	**VetorTag**

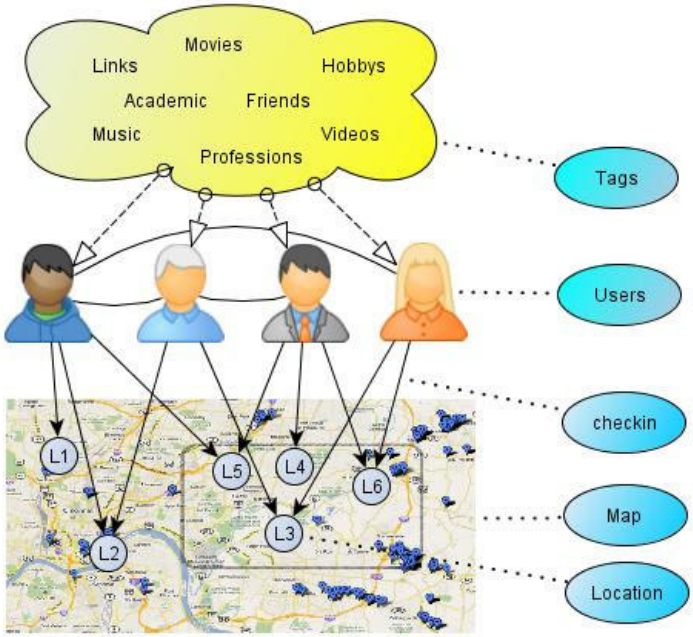

Figura 22: Adaptação e proposta de Recomendação por Tags para a aplicação.

O Capítulo 7 apresentará mais detalhes da arquitetura final da aplicação de recomendação social com alguns testes e resultados para exemplificar a metodologia construída.

7. IMPLEMENTAÇÂO DA APLICAÇÃO DE RECOMENDAÇÃO SOCIAL

A implementação da aplicação com propósito de Recomendação Social do referido Trabalho de Dissertação é apresentada neste capítulo com a descrição das tecnologias utilizadas, a definição da arquitetura voltada para a recomendação social e por fim os experimentos e resultados.

7.1 Tecnologias Utilizadas

7.1.1 *Python, NLTK* e *Numpy*

Python[5] é uma linguagem de altíssimo nível (em inglês, Very High Level Language) orientada a objetos, de tipagem dinâmica e forte, interpretada e interativa. Criada pelo holandês Guido Van Rossum sob o ideal de "Programação de Computadores para todos". Pode ser utilizada tanto na programação em formato Procedural quanto ser adotada ao Paradigma de Orientação a Objetos tendo como principal aspecto positivo uma coleção de módulos e *frameworks* de terceiros que podem ser adicionados. O que contribui na criação de bibliotecas especifica para resolução de determinados problemas em especial a análise de dados.

A Linguagem *Python* é interpretada através de *bytecode* pela máquina virtual do *Python*, tornando o código portável. Com isso é possível compilar aplicações em uma plataforma e rodar em outras ou executar direto do código fonte. É também um software de código aberto com licença compatível com a *General Public License* (GPL), entretanto, permite que o Python seja

[5] Página oficial: http://www.python.org/.

incorporados em produtos proprietários e a especificação da linguagem é mantida pela *Python Software Foundation*[6] (PSF).

Em alguns casos *Python* é utilizado como linguagem script em vários softwares, permitindo automatizar tarefas e adicionar novas funcionalidades. Atualmente a Linguagem é apreciada em várias camadas da indústria e empresas de tecnologia, tais como: *Disney*, *Nokia*, *Google*, *Yahoo*, *Facebook* dispositivos móveis, sistemas de processamento de imagem, entre outros.

Portanto, a utilização desta tecnologia apresentou o potencial necessário para a contemplação da metodologia de recomendação social descrita no capitulo anterior. Tendo como fator decisivo a sua Modularidade e disponibilidade de automatizar determinadas funcionalidades como as técnicas de Mineração de Dados; ideal para a pesquisa de Dissertação.

NLTK[7] (*Natural Language Toolkit*) é uma biblioteca de código aberto que trabalha com Processamento de Linguagem Natural em *Python* visando à análise de textos, sobretudo a Mineração de Dados em Texto. Onde o interpretador executa muitas operações comuns, como tokenização, uso de *tagging* e experimentos de agrupação e classificação de palavras.

Outra tecnologia adotada na pesquisa é o ***Numpy***[8], pacote fundamental para computação científica com *Python*. Ele apresenta, entre outras coisas: um poderoso espaço multidimensional de conjunto de objetos; álgebra linear, transformação Fourier e utilização em números aleatórios; e ferramentas de integração com outras linguagens de programação como C/C++ e Fortran.

Tipos arbitrários de dados podem ser definidos. Isso permite que *NumPy*, de forma transparente e rápida integre-se com uma grande variedade de bancos de dados. *Numpy* está licenciado sob a licença BSD (licença para software livre), permitindo a reutilização com poucas restrições. O *Numpy* tornou-se ideal para utilização em cálculos visando o agrupamento em cima

[6] Endereço na internet da PSF: http://www.python.org/psf/.

[7] Endereço na Internet de NLTK: http://www.nltk.org/book

[8] Página oficial: http://www.numpy.org/

dos resultados dos experimentos com a proposta de uma abordagem Híbrida de Filtragem de Informação na LBSN.

7.1.2 *Application Programming Interface* (API) e JSON

Em geral, os sistemas remotos concedem acessos remotos (web) a seus serviços através de suas API's, dando a possibilidade de incluir a sua funcionalidade em aplicações externas ou nos Web sites. Exemplo de sistemas remotos podem ser as Redes Sociais consolidades como: *Facebook*, *Google+*, *Twitter* e especial para a pesquisa o *Foursquare*. Este recurso contribui para os testes e acesso a dados dos usuários que estão disponibilizadas na Redes Sociais. A **API do *Foursquare*** disponibiliza dados condizentes com seu próprio serviço da LBSN, para que possam ser extraídos e trabalhados por desenvolvedores em formatos como JSON; formato este que pode ser codificado e decodificado na linguagem *Python*.

O **JSON**[9] (*JavaScript Object Notation*) foi desenvolvido utilizando princípios de linguagens C, C + +, C #, *Java*, *JavaScript*, *Perl*, *Python* (apesar de ser independente de linguagem). Portanto, o resultado para os programadores é uma legibilidade e facilidade na analise dos dados inseridos. O formato funciona em uma matriz de valores e objetos que são pares de nome / valor. O nome geralmente vem entre aspas, o valor pode ser uma cadeia entre aspas, ou um número, ou verdadeiro ou falso ou nulo; um objeto ou uma matriz. As estruturas podem ser aninhadas. Por exemplo, abaixo segue uma descrição em formato JSON:

```
{ veiculo:[
{"cor":"azul" , "marca": "Fiat"}
{"cor":"azul" , "marca": "Ferrai"}
{"cor":"amarelo" , "marca": "Chevrolet"}  ]}
```

[9] Página Oficial: http://www.json.org

7.1.3 MYSQL

O **MySQL**[10] é um sistema de gerenciamento de banco de dados relacional (RDBMS) com base em SQL (*Structured Query Language*). MySQL é atualmente disponível sob dois acordos de licenciamento diferentes: gratuitamente, sob a licença GNU *General Public License* (GPL), sistema de código aberto ou através de subscrição do MySQL *Network* para aplicações de negócios. O Banco de Dados MySQL foi adotado como repositório de dados para os experimentos com a aplicação de recomendação social desenvolvida neste trabalho.

7.2 A Ferramenta de Aplicação de Recomendação Social: HYTASO

7.2.1 Caracterização da LBSN – *Foursquare*

Conforme já comentado em capítulo anterior, a Rede Social Baseada em Localização *Foursquare* (criado em 2009) foi adotada para o desenvolvimento da aplicação de Mineração de Dados e, portanto é necessária a descrição e apresentação do seu fluxo de funcionamento. Esta LBSN pode ser acessada tanto em computadores pessoais como dispositivos móveis (celulares, *smartphones* e etc.) e compartilhar a localização do usuário (*check-in)* bastando que o dispositivo apresente um GPS.

Basicamente, no Foursquare (Vasconcelos, 2012), os *check-ins* podem ser realizados em uma variedade de locais (*venues*) e podem ser acumulados na forma de pontos permitindo que os usuários ganhem medalhas (*badges*), prefeituras (*mayorships*) e recebam ofertas do local além de possibilitar o registro de imagens (*upload*) do local. O *venue* é um local no mundo real: uma loja, um hotel, um terminal de ônibus, um aeroporto entre outros. Cada *venue* no Foursquare é classificado em uma das nove categorias pré-definidas: *Arts* &

[10] Página Oficial: http://www.mysql.com

Entertaiment, Colleges, & *Universities, Food, Great Outdoors, Nightlife Spots, Travel Spots, Residences, Professional* & *Other Places* e *Shops* & *Services.* Outra característica do *Foursquare* são as dicas (*tips*) postadas pelos usuários nos *venues*, que refletem suas impressões e comentários positivos ou negativos sobre a visita a esse local. Os check-ins podem ser postados em outras contas de Rede Social como *Facebook e Twitter.*

Ou seja, os *check-ins*, são compartilhados com os amigos e seguidores assim que são postados e têm o diferencial de estarem acessíveis quando se realiza uma busca por algum local nas redondezas ou se acessa um determinado *venue*. Após ler uma *tip*, o usuário pode adicioná-la em sua lista privada de *to-do* ou marcá-la como *done*. O conteúdo de uma lista de *to-dos* é uma informação privada ao usuário e sua rede social. O número de vezes que uma *tip* foi marcada como *done* é uma informacão disponível no *Foursquare* e serve com estimativa da quantidade de *feedback* vinda de outros usuários que leram a *tip*, além de ser uma mecanismo de identificar boas recomendações a serem seguidas (Vasconcelos, 2012).

Os usuários também podem ganhar distintivos: para *check-in* em locais com determinadas *tags*, para frequentes *check-ins*, ou outros padrões, como o tempo aproximado de um *check-in.* A LBSN divide o usuário em três categorias: *users, celebrities e brands.* A diferença principal entre eles é um grau de relacionamento que existe entre outros usuários. O *user* é o mais comum só pode relacionamentos de amizade tipo mútua ou seguir usuários de outros tipos. O usuário do tipo *celebrity* possui os dois tipos de relacionamento: amizade e seguidor/seguido. E por fim, os usuários do tipo *brand* são usuários encarregados de postar conteúdo na forma de *tips*, fotos e comentários e só podem se relacionar com seguidores.

O conteúdo e as relações sociais existentes nos componentes do *Foursquare* (os comentários *tips*, as listas *to-dos*, os *users* e outras páginas da LBSN) citados anteriormente **estão inseridos no mecanismo de Recomendação por *Tags***. As Figuras 23, 24, 25 e 26 ilustram a interface do *Foursquare* em um dispositivo móvel. As Figuras 27 e 28 ilustram as interfaces de busca feita pelo Usuário.

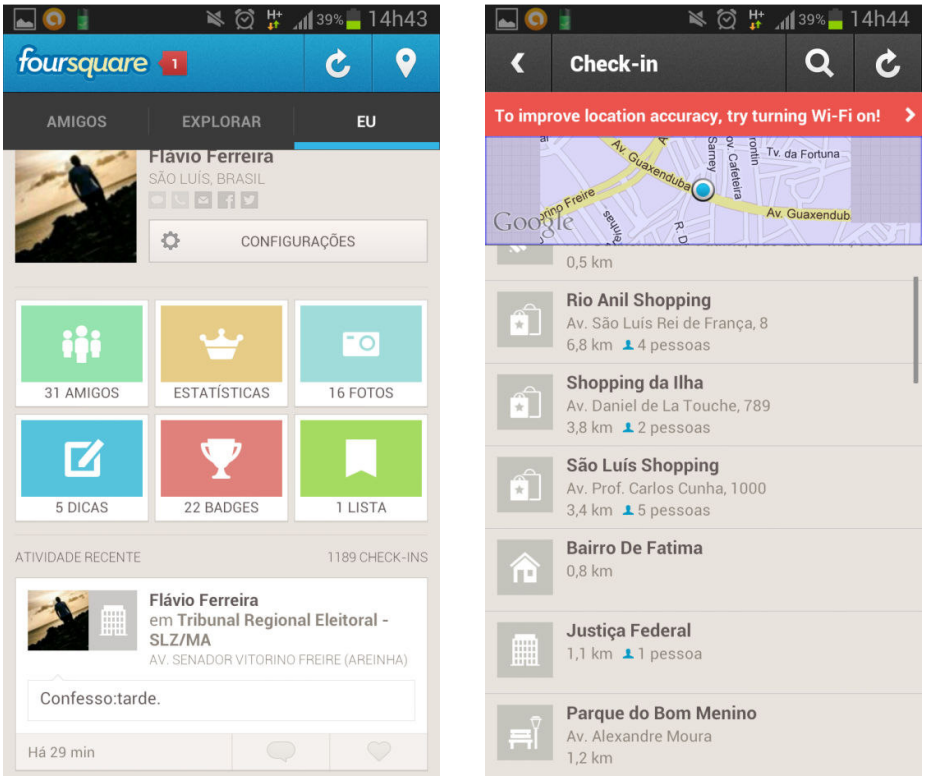

Figura 23: a) Página Inicial do *Foursquare.* b) Página Inicial do *Check-in*.

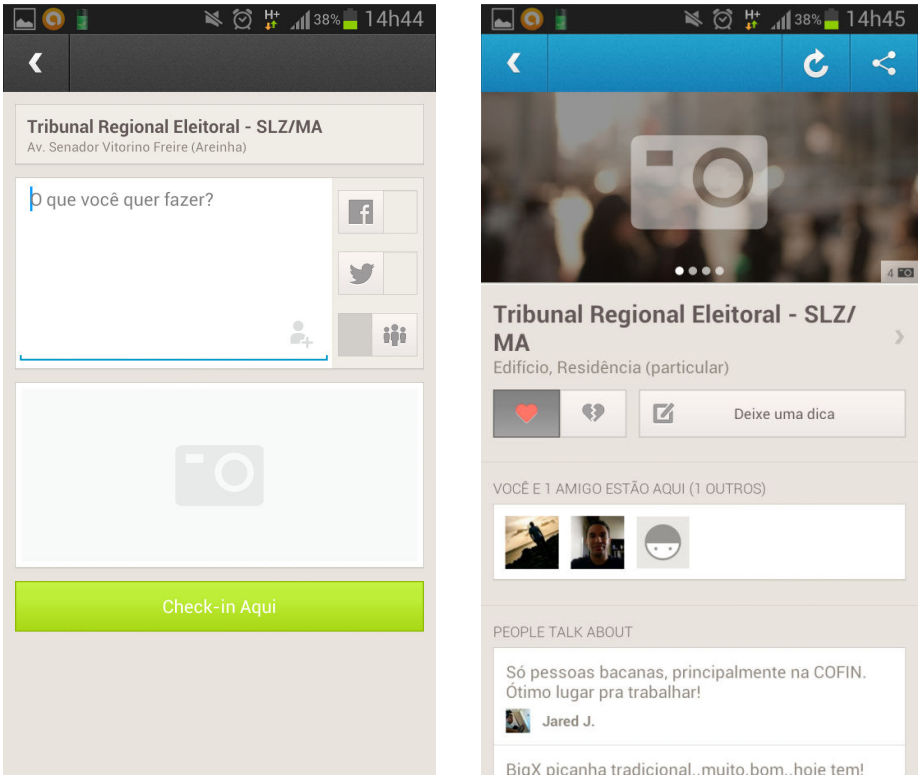

Figura 24: a) Efetuação do *check-in.* b) Página de interação do *Check-in*.

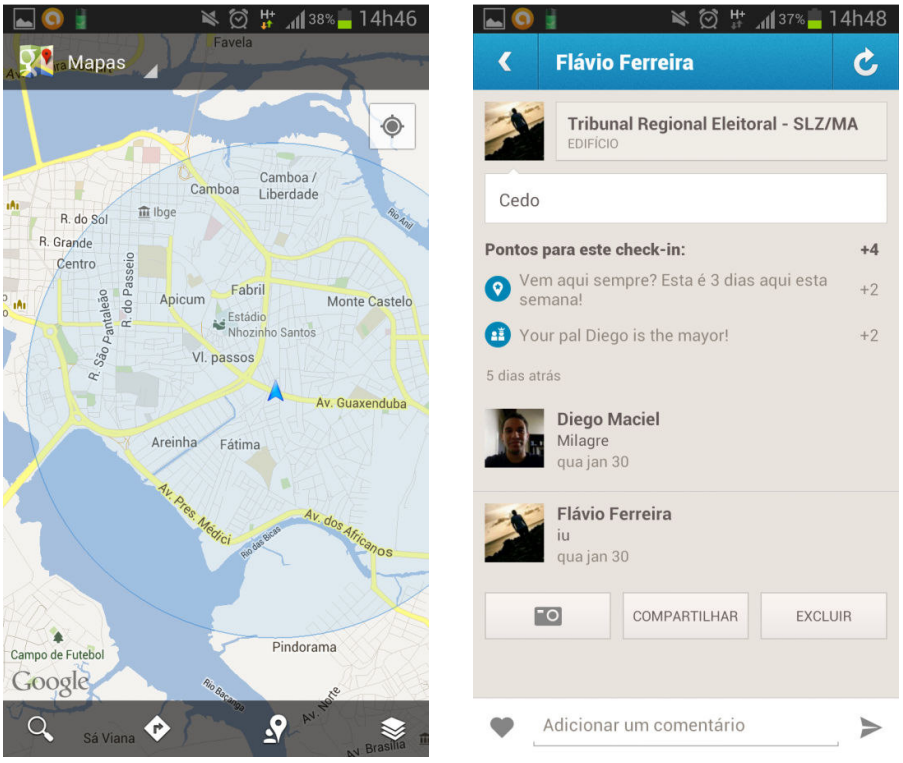

Figura 25: a) Mapa na LBSN. b) Página com comentários do *Check-in*.

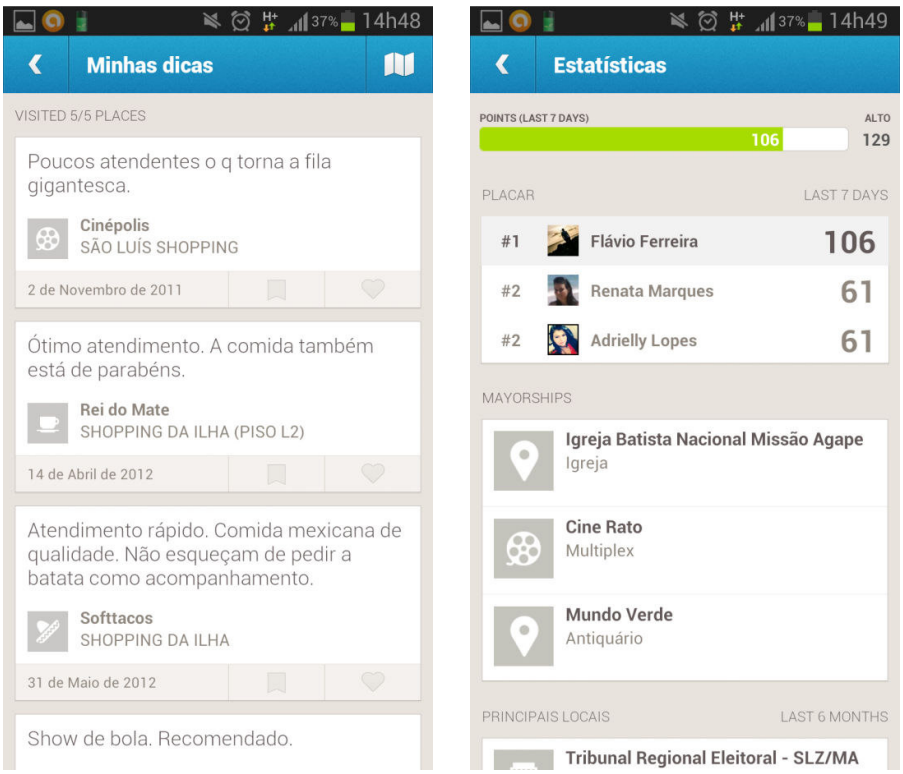

Figura 26: a) Lista de *Tips* (dicas). b) Registro de Estatísticas do Usuário.

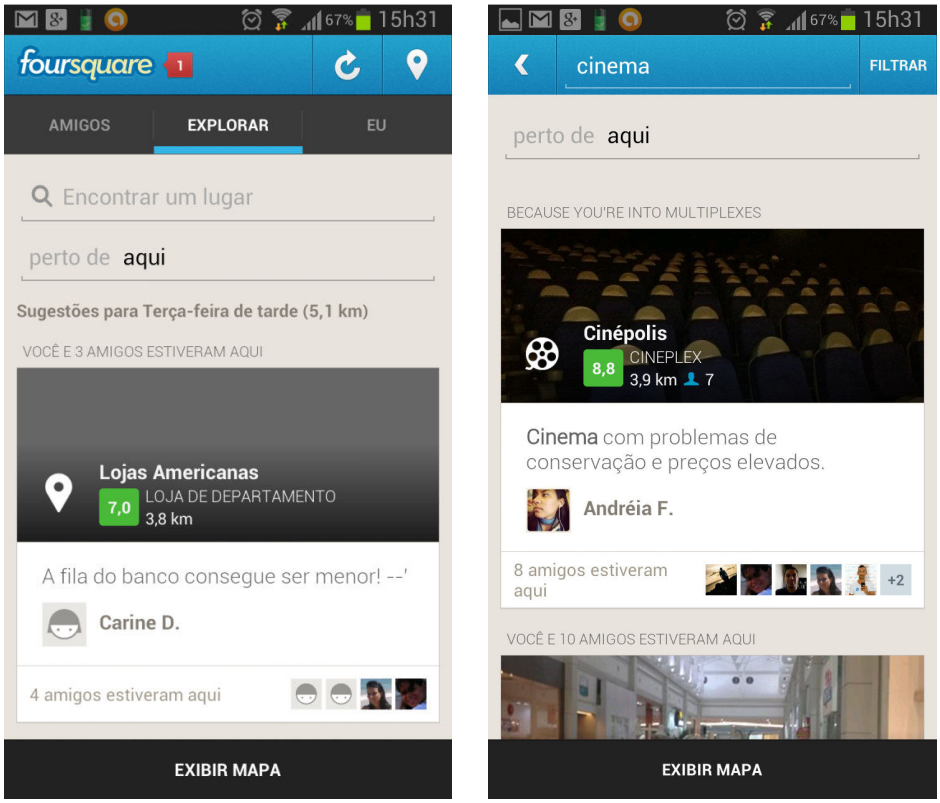

Figura 27: Telas de busca de recomendações "Explorar" no *Foursquare*.

Figura 28: Interface de busca de recomendações "Pesquisar" no *Foursquare*.

7.2.2 Arquitetura e Modelagem da Aplicação HYTASO

Nesta pesquisa é proposto uma abordagem Híbrida para a filtragem dos dados na LBSN *Foursquare*, e depois, através das técnicas Mineração de Dados Textuais, realizar a Recomendação Social com o uso das Tags. Esta aplicação foi intitulada com o nome de **HYTASO** sigla dos termos *Hybrid Tagging Social*. A Figura 29 ilustra a metodologia da pesquisa que contempla a criação de uma arquitetura dividida nas seguintes etapas:

1. Inserção das Credenciais do Usuário na Rede Social *Foursquare*;
2. Consulta de termos feitos pelo usuário;
3. Coleta de dados (*Crawler*) junto aos parâmetros próprios do *Foursquare*;
4. Técnica de Mineração de Dados em Textos visando a extração das componentes da LBSN;
5. Aplicar a Recomendação por *Tags* direcionada aos componentes do *Foursquare*;
6. A Recomendação Social com os dados encontrados retornados para o usuário.

Inicialmente devem-se inserir as credenciais de acesso aos dados da Rede Social. Este trabalho tratou de criar um mecanismo automático de gerador das credenciais: ID e código de acesso. A obtenção e manipulação dos dados de um perfil somente são possíveis através de uma API de desenvolvimento onde existe uma base de dados web com as informações dos usuários na rede social.

Esta base de dados do Foursquare está distribuída através de dados desestruturados em formato JSON (*JavaScript Object Notation*). Ou seja, qualquer informação, compartilhamento, comentário ou atualização no perfil do usuário é discriminado em uma estrutura JSON pronto para ser utilizado no desenvolvimento de aplicações futuras.

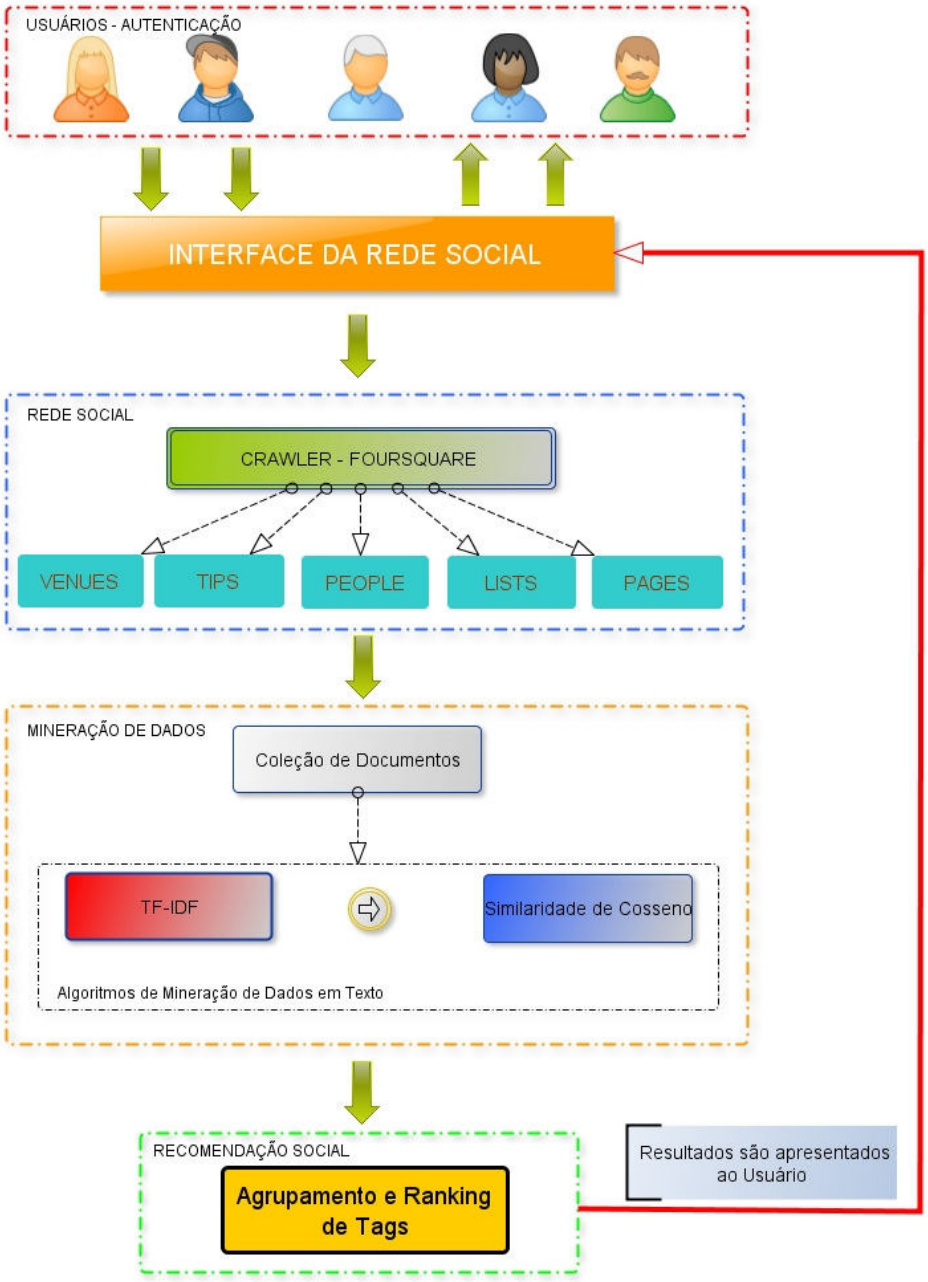

Figura 29: Arquitetura em camadas da Aplicação de Recomendação Social HYTASO.

Uma vez que os dados estão dispostos em entidades no formato JSON da própria rede social, foi necessária a criação de um mecanismo de extração de dados único que dialogasse com a estrutura: o *Crawler*. O mecanismo *crawler* coleta as informações referenciadas nas entidades da estrutura JSON, onde as entidades também servem como parâmetros no ato da consulta

realizada pelo usuário. Junto aos dados das entidades que são coletadas há também dados geolocalizados incorporados como: latitude, longitude, altitude, cidade e etc., que ajudam no processo de mapeamento de padrões de informação em determinados locais. A Figura 30 apresenta a entidade *venue* (formato JSON) e junto a sua estrutura algumas informações de geolocalização.


```
venue: {
    id: "4ff764c3e4b01ead023da588",
    name: "Brooklyn Beach Shack",
    contact: { },
    location: {
        lat: 40.700051,
        lng: -73.996479,
        distance: 297,
        postalCode: "11201",
        city: "Brooklyn",
        state: "NY",
        country: "United States",
        cc: "US"
    },
```


Figura 30: Estrutura da entidade do *venue* no formato JSON.

As técnicas de Mineração de Dados, relatadas no Capítulo 6, utilizarão os dados coletados pelo *crawler* a fim de extrair os resultados próximo do que necessita ser recomendado. As pontuações geradas delimitam os resultados mais relevantes para a pesquisa do usuário e permite a construção de uma fundamentação semântica nas *tags* – tendo em vista que a forma que se estruturam e se relacionam as entidades do *Foursquare* as caracterizam como *tags.* A estruturação dos dados em forma de uma hierarquia de entidades aponta para a co-ocorrência de *tags* (*colaboration tagging*); inferindo, por exemplo, que *tags* que co-ocorrem freqüentemente em um mesmo recurso sugerem uma similaridade semântica.

Em seguida é ilustrado a Modelagem UML da Aplicação HYTASO com as Figuras 31 e 32 representando os Diagramas de Atividade, e a Figura 33 ilustrando o Diagrama de Classes com toda a estrutura da Aplicação

Embarcada do trabalho de Dissertação. Cada classe apresenta nome referente com a função desempenhada na aplicação

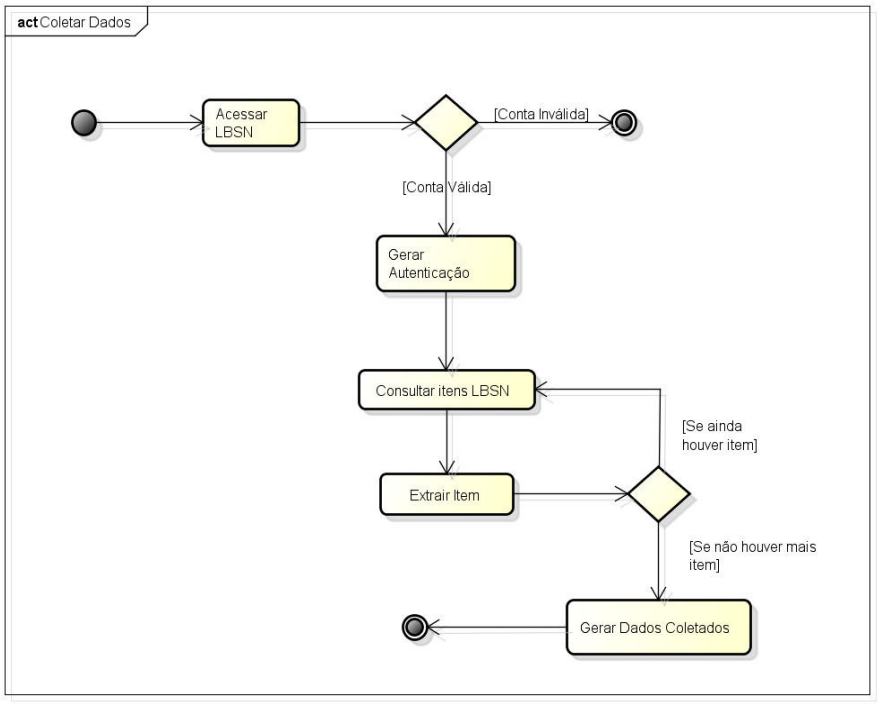

Figura 31: Representação do Diagrama de Atividade Coletar Dados.

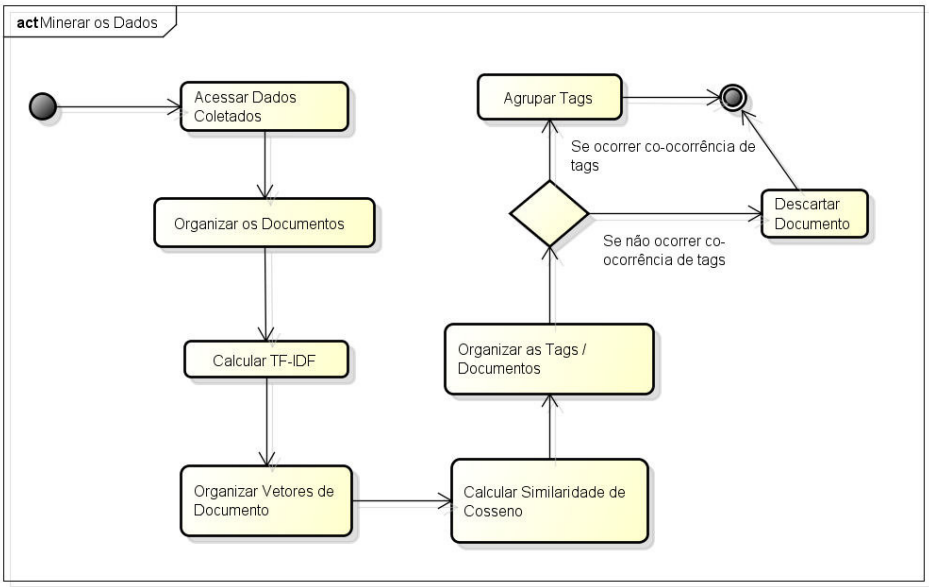

Figura 32: Representação do Diagrama de Atividade Minerar Dados.

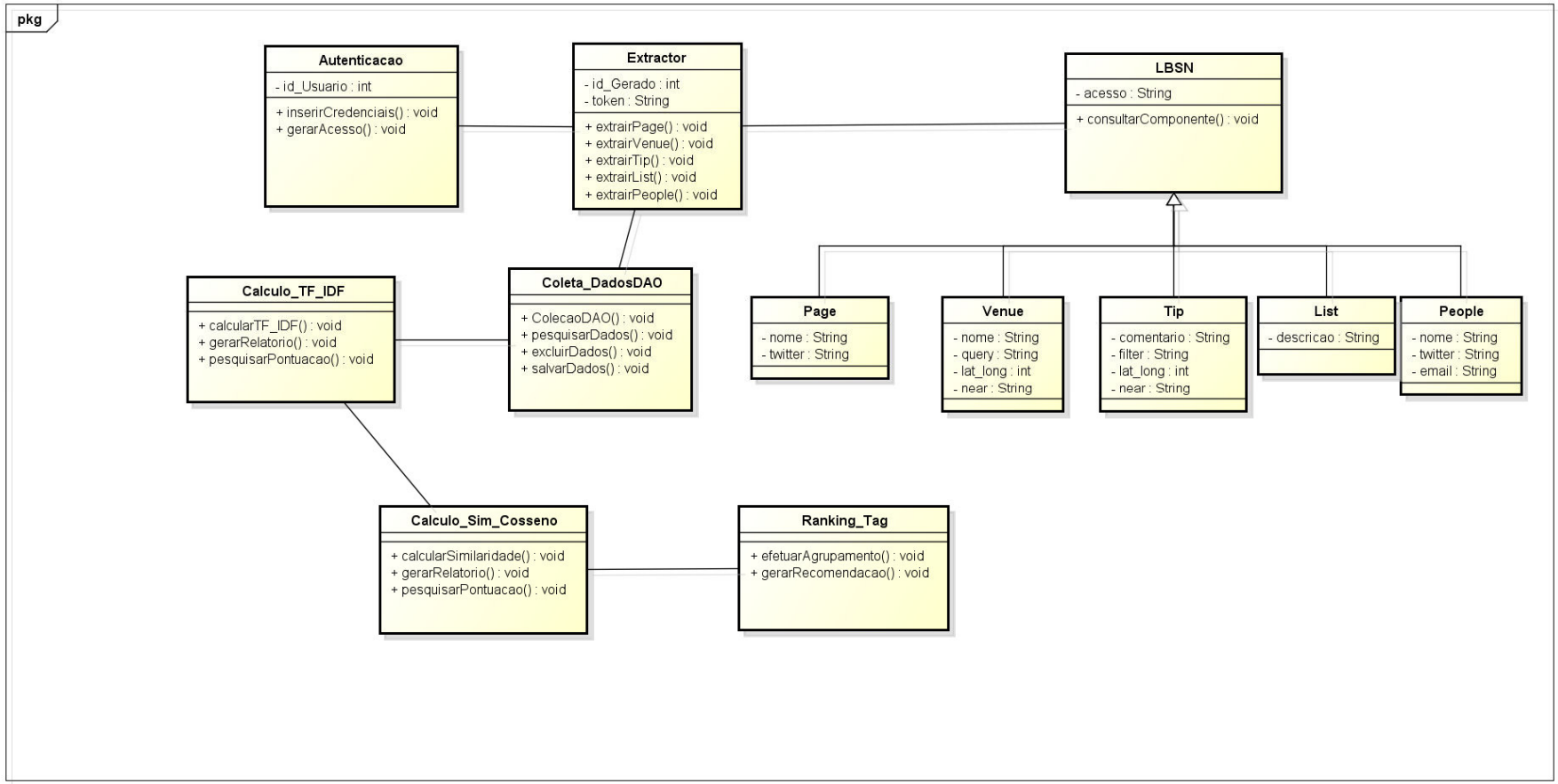

Figura 33: Representação em UML do Diagrama de Classes.

A classe **Autenticação** fica responsável por gerar as credenciais necessárias para que a classe ***Extractor*** possa coletar junto a Rede Social Baseada em Localização do *Foursquare* as informações textuais que estão contidas em seus componentes. Para estes também foram construídas classes (***Pages***, ***Venue***, ***Tip***, ***List*** e ***People***) que trabalham através de parâmetros que condizem com a busca do usuário, tais como: palavras-chaves, nomes de lugares ou ainda nomes de pessoas.

A coleta dos Dados é feita através de parâmetros existentes em cada tipo de componente pertencente ao *Foursquare* como, por exemplo, *Pages e People são* componentes que apresentam um parâmetro denominado *query* que pode ser passado qualquer palavra ou enunciado para que seja bsucado na LBSN. Os nomes das componentes coincidem com os nomes das classes para garantir um fácil entendimento do que se refere o código.

A Classe ***Extractor*** é onde está posicionado os *crawlers* responsáveis por extrair todos os dados e informações textuais pertinentes a cada componente da rede social. Este tipo de recurso só é possível com os dados credenciais gerados na classe **Autentificação**, para qualquer credencial passada incorretamente a coleta não será realizada.

Os dados textuais coletados são armazenadas em um repositório de dados (através da classe **ColetaDAO**) para que a classe **Calculo_TF_IDF**, responsável por trabalhar com os métodos do algoritmo TF-IDF (*Term Frequency Inverse Document Frequency*), execute seu raciocínio e gere as pontuações esperadas em cima de cada corpus de documento. A Figura 34 apresenta o método na Linguagem *Python* responsável pelo cálculo TF-IDF.

```
def tf_idf(termo, documento, corpus):
    return tf(termo, documento) * idf(termo, corpus)
```

Figura 34: Representação do método TF-IDF em *Python*.

As pontuações geradas junto ao corpus de documento são formalizados para cada corpus de documento de modo que sejam divididos em forma de “vetores” com índices contendo as pontuações para cada corpus de documento, onde a classe **Calculo_Sim_Cosseno** utiliza de seus métodos para realização do método de Similaridade de Cosseno. A Figura 35 apresenta o método responsável pelo caçulo da Similaridade de Cosseno em Linguagem *Python*.

```
distancia = nltk.cluster.util.cosine_distance
(vetor_documento1,vetor_documento2)
```

Figura 35: Representação do método Similaridade de Cosseno em *Python*.

É importante ressaltar que o repositório de dados textuais na classe **ColetaDAO** armazena e dispõe os resultados para que possam ainda ser utilizados em outras técnicas.

Por último, a classe ***Ranking_Tag*** realiza o agrupamento e ordenação dos corpus de documentos baseado no grau de similaridade existente na colaboração entre as *tags* (*tagging social*). Ao final é gerado a recomendação das *tags* existentes nos corpus de documento.

Os Diagramas de Atividade representados anteriormente apresentam os principais eventos de mudança de estados quanto a aplicação desenvolvida. São dois processos distintos, entretanto o desenvolvimento da atividade de **Minerar Dados** tem início após a coleta adequada dos dados feita na atividade **Coletar Dados**.

Nos Apêndices deste Trabalho de Dissertação é apresentado a principal codificação criada para o desenvolvimento e funcionamento da aplicação.

7.3 Experimentos e Resultados

Para demonstrar a Aplicação HYTASO de Recomendação Social criada foram elaborados alguns testes com a metodologia desenvolvida em um cenário de consulta na LBSN do *Foursquare*.

Inicialmente, o processo de consulta dependerá das credenciais que caracterizam como um usuário da LBSN. Conforme comentado na seção anterior foi criado um mecanismo que gera automaticamente o ID e *token* de acesso aos dados que se encontram disponíveis na API de Desenvolvimento da Rede Social Baseada em Localização para serem trabalhados na pesquisa. Abaixo é apresentada a estrutura do método que permite acessar recursos do *Foursquare* através de URL na API de desenvolvimento:

- *https://api.foursquare.com/v2/"componente do Foursquare"/search?"parâmetro do Foursquare"="token de acesso"="id de acesso"*

Onde:

- Componente do *Foursquare* se divide em: *user*, *venue*, *tip*, *page* e *list*. Cada um destes retorna os dados pertinentes ao seu domínio. Por exemplo, *user* contém os dados relevantes aos usuários da rede social.

- Parâmetro do *Foursquare* está condizendo com o tipo de componente escolhido. Por exemplo, o componente *venue* apresenta o parâmetro "*query*" onde cataloga palavras-chave.

- *Token* de Acesso é o código gerado na autenticação do usuário, sendo restrito a cada usuário. O código apresenta um conjunto de caracteres em letras e números. Por exemplo: SXMHXJI2BIW1TKRXOBHM33

- *Id* de Acesso compõem por último a autenticação para formalizar o acesso a coleta de dados sendo formado por números apenas.

O *Crawler* construído procurou montar uma Base de Dados com resultados pertinentes a consulta do usuário que agregasse informações relativas ao Componente do *Foursquare*.

A pesquisa de Dissertação utilizou como experimento a consulta por termos referente às palavras “caranguejo” e “Coldplay”. Onde para cada uma foram extraídos aproximadamente 100 corpus de documentos – a escolha desta quantidade se deve ao número limite de resultados retornados pela LBSN. Cada corpus apresenta desde nomes comuns a frases, como por exemplo: “Taylor Swift and Paula Fernandes music”. Estes corpus refletem nomes de usuários (*users*); locais de visitação (*venues*); comentários de usuários ou de amigos do usuário (*tips*); listas de coisas a se fazer (*To-dos*); e páginas de usuários ou de locais ou ainda de itens (*pages*).

Após a coleta de dados, o algoritmo TF-IDF calculará a freqüência dos termos que se apresentam nos corpus identificando os termos mais próximos daquilo que o usuário deseja (de acordo com a metodologia desenvolvida e apresentada no Capítulo 6), tendo em vista a dificuldade que foi encontrada em se recomendar algo similar ao que usuário procura, seja em uma busca direta feita por ele ou sugestões dadas pela própria LBSN na tentativa de determinar um padrão de comportamento do usuário.

Portanto, o primeiro experimento apresentou os seguintes termos de consulta: “**caranguejo”**; “**torta de caranguejo”**; “**restaurante caranguejo”**. Este exemplo foi adotado tendo como quadro de localidade da busca o Estado do Maranhão. E apresenta um grau de dificuldade em encontrar resultados que sejam pertinentes ao que usuário deseja realmente encontrar.

Cada corpus de documento remete aos valores de uma entidade JSON pertencente a LBSN. Por exemplo, o corpus C1 apresenta o corpus: "Camaroada, arroz de cuxa, torta de caranguejo... Otimo restaurante maranhense..."; portanto cada corpus apresenta dados textuais de locais, comentários ou ainda informações pertinentes a pessoas e itens. Abaixo segue a descrição de alguns dos corpus de documentos extraídos pelo *Crawler*:

- "Melhor caranguejo e cerveja beeem gelada..."
- "Caranguejo no leite de coco delicioso! O arroz de toucinho e o melhor da cidade."
- "A coxinha de caranguejo do Ferreiro Grill e matadora. Faz o maior sucesso em Aracaju. Sera que no de Sao Luis segue o mesmo padrao?"
- "Caranguejo no leite de coco delicioso! O arroz de toucinho e o melhor da cidade."
- "O melhor caranguejo atendimento perfeito"
- "Evite a torta de caranguejo congelada"
- "Restaurante Cabana do Sol"

A Tabela 12 e Tabela 13 apresentam as pontuações cumulativas em cima de cada corpus de documento através da utilização do cálculo da freqüência dos termos com o TF-IDF. Os resultados podem ser contabilizados como a soma de valores para os termos que aparecem no mesmo corpus. A pontuação acumulativa gera um resultado mais expressivo para a consulta, mesmo não considerando a proximidade ou ordenação das palavras em um corpus. Algumas pontuações que apresentam predomínio de *stopwords* não são apresentadas na tabela. As pontuações dos três termos de consulta foram organizadas nas tabelas para cada corpus de documento, essa formalização ajudará na próxima etapa da Mineração de Dados em Texto dos documentos da LBSN. As Tabelas 12 e 13 ilustram os corpus dos documentos existentes no *Foursquare* em forma código: A16, C1, E1 e etc.

Cada código representa um comentário feito na LBSN, uma indicação de localidade, a página de algum perfil de usuário ou empresa, ou seja, a codificação representa qualquer descrição textual condizente com os dados solicitados inicialmente na consulta (na seção do Anexo serão apresentadas algumas destas descrições). A organização dos caracteres do código não apresenta nada de especial, sendo organizado em pares de letra e número como forma de facilitar na programação dos algoritmos.

Tabela 12: Primeira Relação de Pontuações para cada termo de consulta.

Corpus	Pontuação	Corpus	Pontuação	Corpus	Pontuação
A1	0,766; 1,314; 0,766	**A19**	0,209; 0,508; 0,209	**C1**	0,000; 0,727; 0,210
A2	0,000; 0,000; 0,000	**A20**	0,383; 0,383; 0,383	**C2**	0,383; 1,321; 0,383
A3	0,000; 0,000; 0,000	**A21**	0,383; 0,383; 0,383	**C3**	0,000; 0,000; 0,000
A4	0,000; 0,000; 0,000	**A22**	0,460; 0,460; 0,460	**C4**	0,000; 0,329; 0,000
A5	0,000; 0,000; 0,000	**A23**	0,074; 0,233; 0,074	**C5**	0,000; 0,329; 0,000
A6	1,150; 1,150; 1,150	**A24**	0,383; 0,383; 0,383	**C6**	0,000; 0,548; 0,000
A7	0,092; 0,223; 0,092	**A25**	0,153; 0,482; 0,153	**C7**	0,000; 0,329; 0,000
A8	0,000; 0,727; 0,210	**A26**	0,383; 1,321; 0,383	**C8**	0,000; 0,548; 0,000
A9	0,000; 0,063; 0,000	**A27**	0,000; 0,235; 0,000	**C9**	0,000; 0,164; 0,000
A10	0,115; 0,279; 0,115	**A28**	0,000; 0,000; 0,000	**C10**	0,000; 0,183; 0,000
A11	0,328; 0,563; 0,328	**A29**	0,192; 0,466; 0,192	**C11**	0,000; 0,411; 0,000
A12	0,192; 0,329; 0,192	**A30**	0,328; 0,563; 0,328	**C12**	0,000; 0,411; 0,000
A13	0,153; 0,263; 0,153	**A31**	0,177; 0,430; 0,177	**C13**	0,000; 0,274; 0,000
A14	0,287; 0,287; 0,287	**A32**	0,000; 0,000; 0,000	**C14**	0,000; 0,329; 0,000
A15	0,153; 0,263; 0,153	**A33**	0,766; 0,766; 0,766	**C15**	0,000; 0,658; 0,000
A16	0,096; 0,301; 0,096	**A34**	0,000; 0,000; 0,000	**C16**	0,000; 0,548; 0,000
A17	0,177; 0,177; 0,177	**A35**	1,150; 1,150; 1,150	**C17**	0,000; 0,411; 0,000
A18	0,460; 0,789; 0,460	**A36**	0,766; 1,314; 0,766	**C18**	0,000; 0,274; 0,000

Tabela 13: Segunda Relação de Pontuações para cada termo de consulta.

Corpus	Pontuação	Corpus	Pontuação	Corpus	Pontuação
C19	0,000; 0,411; 0,000	**E4**	0,000; 0,000; 0,525	**E21**	0,000; 0,000; 0,700
C20	0,000; 0,548; 0,000	**E5**	0,000; 0,000; 0,700	**E22**	0,000; 0,000; 1,049
C21	0,000; 0,000; 0,000	**E6**	0,000; 0,000; 0,700	**E23**	0,000; 0,000; 0,700
C22	0,000; 0,411; 0,000	**E7**	0,000; 0,000; 0,700	**E24**	0,000; 0,000; 0,700
C23	0,000; 0,164; 0,000	**E8**	0,000; 0,000; 0,700	**E25**	0,000; 0,411; 0,525
C24	0,000; 0,205; 0,000	**E9**	0,000; 0,000; 0,525	**E26**	0,000; 0,000; 1,049
C25	0,000; 0,329; 0,000	**E10**	0,000; 0,000; 1,049	**E27**	0,000; 0,000; 1,049
C26	0,000; 0,329; 0,000	**E11**	0,000; 0,000; 1,049	**E28**	0,000; 0,000; 0,525
C27	0,000; 0,274; 0,000	**E12**	0,000; 0,000; 0,525	**E29**	0,000; 0,000; 0,700
C28	0,000; 0,000; 0,000	**E13**	0,000; 0,000; 0,350	**E30**	0,000; 0,000; 0,525
C29	0,000; 0,205; 0,000	**E14**	0,000; 0,000; 0,700	**E31**	0,000; 0,000; 0,525
C30	0,000; 0,411; 0,000	**E15**	0,000; 0,000; 0,700		
C31	0,000; 0,329; 0,000	**E16**	0,000; 0,000; 0,525		
C32	0,000; 0,000; 0,000	**E17**	0,000; 0,000; 0,525		
E1	0,000; 0,727; 0,210	**E18**	0,000; 0,000; 2,099		
E2	0,000; 0,000; 1,049	**E19**	0,000; 0,000; 0,700		
E3	0,000; 0,000; 0,350	**E20**	0,000; 0,411; 0,525		

A segunda etapa das técnicas de Mineração de Dados em Texto é a utilização do algoritmo de Similaridade de Cosseno como forma de comparar cada vetor de documento formalizado após o uso do TF-IDF. A Tabela 14 apresenta alguns resultados com a aplicação da métrica de similaridade na comparação de um corpus com outro corpus da coleção de documento através das três pontuações dos termos de consulta presente em cada corpus de documento.

O resultado final compreende um total de 4.851 comparações de graus de similaridade entre vetores de documentos (conforme a seção de Similaridade de Cosseno descrito na seção do Capítulo 6). Portanto, a tabela abaixo apresenta “pesos” referentes às comparações de um vetor de documento com os demais vetores de documento restante da coleção.

Tabela 14: Relação de alguns Pesos comparativos entre vetores de documentos.

Vetores	Graus	Vetores	Graus	Vetores	Graus
A25 x A26	0.00053879	**A25 x C1**	0.04317	**A25 x E1**	0.71041
A25 x A27	0.08770	**A25 x C2**	0.00053	**A25 x E2**	0.71041
A25 x A28	Nan	**A25 x C3**	Nan	**A25 x E3**	0.710413
A25 x A29	0.00557	**A25 x C4**	0.08770	**A25 x E4**	0.710413
A25 x A30	0.03548	**A25 x C5**	0.08770	**A25 x E5**	0.710413
A25 x A31	0.005531	**A25 x C6**	0.08770	**A25 x E6**	0.710413
A25 x A32	NaN	**A25 x C7**	0.08770	**A25 x E7**	0.710413

As pontuações onde o resultado se intitula “Nan” significam dados que não puderam ser obtidos devido a um dos vetores de documento apresentarem em todas as suas pontuações dados nulos sendo assim insuficiente a aplicação da métrica de similaridade.

O algoritmo TF-IDF cria pontuações aos documentos encontrados através de uma Filtragem Baseada em Conteúdo resultando em uma classificação dos documentos mais relevantes ao que o usuário busca. Com as pontuações geradas anteriormente, o algoritmo de Similaridade de Cosseno conduz a Filtragem Baseada na Colaboração com o objetivo de traçar a similaridade entre os documentos encontrados e com a consulta feita pelo usuário. A abordagem Híbrida combina estas duas técnicas de Filtragem de Informação (em outras pesquisas ambas são utilizadas em casos isolados), os documentos recuperados são formalizados para que sejam adotados posteriormente na Recomendação por *Tags*. A terceira etapa é a utilização do algoritmo do Vizinho mais Próximo (k-NN) sobre as pontuações dos vetores de documento como forma de agrupar as *tags*. A Tabela 15 apresenta somente uma parte dos documentos em *ranking*, pois todo conteúdo analisado é bem amplo para ser detalhado abaixo.

Tabela 15: Relação de alguns corpus de documentos de agrupados.

Corpus de Documento	**Pontuação para *ranking***					
	Tag 2	**Tag 3**	**Tag 4**	**Tag 5**	**Tag 6**	**Tag N**
Tag 1	0.8204	0.7235	0.7104	0.6446	0.4226	**...**
	Tag 1	**Tag 3**	**Tag 4**	**Tag 5**	**Tag 6**	**Tag N**
Tag 2	0.9015	0.7348	0.5723	0.5541	0.7150	**...**
	Tag 1	**Tag 2**	**Tag 4**	**Tag 5**	**Tag 6**	**Tag N**
Tag 3	0.5245	0.7104	0.4721	0.5547	0.8532	**...**
	Tag 1	**Tag 2**	**Tag 3**	**Tag 5**	**Tag 6**	**Tag N**
Tag 4	0.9345	0.8235	0.8204	0.6446	0.5456	**...**
	Tag 1	**Tag 2**	**Tag 3**	**Tag 4**	**Tag 6**	**Tag N**
Tag 5	0.4830	0.7132	0.7104	0.6340	0.4234	**...**

A tabela anterior mostra às pontuações da análise de similaridade sucessivas de uma *tag* a outra *tag,* constituída nos corpus dos documentos de forma que construa o vetor de *tags* delimitado no capítulo 6 (Cattuto, 2008). Os melhores resultados serão utilizados na recomendação social ao usuário. Tomando como exemplo as pontuações iniciais das *tags*, o Vetor de *Tag* 3 (0.5245; 0.7104; 0.4721; 0.5547; 0.8532;...) apresenta pesos menos expressivos na similaridade em comparação com as demais *Tags*, enquanto o Vetor de *Tag* 2 com seus resultados (0.9015; 0.7348; 0.5723; 0.5541; 0.7150;...) demonstra ter uma similaridade bem mais alta com as *Tags* que consigo são analisadas, sendo justificado na seguinte conclusão: os pesos de maior valor encontrados onde há a proximidade do valor 1 denota uma maior precisão na similaridade dos documentos analisados, enquanto pesos próximos do valor 0 representam uma baixa similaridade entre os documentos. A abordagem de Recomendação por Tags vem complementar as abordagens de Filtragem de Informação Baseada em Conteúdo e Filtragem de Informação Baseada na Colaboração. A própria colaboração dos usuários de uma LBSN, como o *Foursquare*, é o fator único para recomendação social, esta combinação de técnicas procura suprir uma carência existente na Rede Social.

Com base nos resultados encontrados nos corpus de documentos e considerando que cada corpus de documento é intitulado como uma *Tag* através do uso de entidades JSON (conforme Figura 28) na LBSN adotada; é possível a descoberta de dados geolocalizados relativos aos documentos, conforme destacado no Capítulo 6. A hierarquia das entidades JSON permite extrair os relacionamentos entre *tags* a partir da *tag* que é sugerida ao usuário (como resultado da pesquisa dos termos de consulta). Tomando como exemplo o termo de consulta “caranguejo”, obtendo um corpus de documento como de um comentário no *Foursquare* (*tip*) como recomendação social; pode se extrair com o relacionamento das *tags:*

- A *Tag* do perfil do usuário (o usuário que comentou);
- A *Tag* do local onde foi feito o comentário;

- E através do perfil do usuário que comentou obter a *Tag* do local onde vive o usuário, assim como através da identificação do perfil do usuário as *Tags* das listas de locais mais freqüentados por ele.

A Figura 36 e Figura 37 apresentam trechos das entidades JSON de um comentário sugerido ao termo de consulta “caranguejo” como recomendação a determinado usuário onde é possível obter dados de localização e registros do usuário que fez o comentário.


```
{
   id: "4f9038f9e4b027bf9d29f5af",
   createdAt: 1334851833,
   text: "Camaroada, arroz de cuxã, torta de caranguejo... Otimo
   restaurante maranhense...",
   canonicalUrl: "https://foursquare.com/item/4f9038f9e4b027bf9d29f5af",
   likes: {
      count: 0,
      groups: [ ]
   },
   like: false,
   logView: true,
   todo: {
      count: 0
   },
   venue: {
      id: "4f2bee6ce4b03d6f92930ca2",
      name: "Ilha dos Sabores",
      contact: { },
      location: {
         lat: -2.5272974877421706,
         lng: -44.25511729185479,
         distance: 3372,
         country: "Brazil",
         cc: "BR"
      },
      canonicalUrl: "https://foursquare.com/v/ilha-dos-
      sabores/4f2bee6ce4b03d6f92930ca2",
```


Figura 36: Entidades JSON com dados geolocalizados nos valores de das *tags* associadas.

A análise de *tags* permite descobrir as preferências de um determinado usuário e fornecer sugestões para o usuário de itens que poderiam ser interessantes. A vantagem da Recomendação por *Tags* é que as preferências do usuário e interesses são expressas por *tags* usadas por determinada pessoa. Isso contribui na formação de recomendações sociais mais precisas e personalizadas.

```
    user: {
        id: "21271720",
        firstName: "Viviane",
        lastName: "F.",
        gender: "female",
        photo: {
            prefix: "https://irs0.4sqi.net/img/user/",
            suffix: "/BUJ1NUSPYCHQJANO.jpg"
        }
    }
},
```


Figura 37: Entidade JSON *User* da *tag do* usuário que fez o comentário.

De acordo com a implementação teste realizada, considera-se que esta metodologia desenvolvida rompe com alguns problemas Filtragem de Informação: a Superespecialização nos Sistemas de Recomendação e a limitação do *Cold Start.*

De acordo com a implementação teste realizada, considera-se que esta metodologia desenvolvida rompe o problema da Superespecialização nos Sistemas de Recomendação, pois uma das vantagens em trabalhar com a recomendação social por *tags* é a agilidade e liberdade em definir marcações semânticas (Folksonomia) e não necessitar de uma aprovação para acrescentar novos termos, tirando a exclusividade de recomendar aos usuários apenas itens que já foram avaliados por eles.

A limitação do *Cold Start Problem* pode ser diminuída pela utilização da colaboração de *tags*, após a atualização automática das informações compartilhadas há uma rápida propagação dos recursos pela Rede Social chegando até mesmo aos usuários a recomendação de itens ainda recentes. O diferencial é a possibilidade de o usuário categorizar os termos em um vocabulário próprio e propagá-lo de forma indireta.

8. CONSIDERAÇÕES FINAIS

Conforme em outros capítulos da Dissertação, a proposta da pesquisa é apresentar uma aplicação para efetuar a Recomendação Social através de uma abordagem Híbrida para Filtragem de Informação utilizando técnicas de Mineração de Dados em Texto em Redes Sociais Baseadas em Localização. O trabalho apresenta algumas soluções para algumas deficiências encontradas nas LBSN's (decorrentes das abordagens dos Sistemas de Recomendação). A pesquisa contempla ainda um estudo a respeito da estrutura dos dados da Rede Social em entidades (estruturados no formato JSON) e a concepção de um rastreador para a coleta de dados que pode ainda ser utilizado em outras abordagens de recomendação que necessitem extrair os dados de uma Rede Social.

8.1 Contribuição do trabalho

Uma das contribuições da pesquisa diz respeito ao estudo das *Tags* e seu potencial semântico, pois, permitem aos usuários a criação de um vocabulário próprio e propagação da descoberta de conhecimento em áreas inexploradas.

A utilização do algoritmo TF-IDF procurou através das pontuações e os somatórios cumulativos aproximarem a similaridade entre os itens (locais, hábitos, lista de itens, comentários), ou outros perfis procurados no ato da consulta do usuário. A técnica busca reduzir as ambigüidades e redundâncias que são encontradas nas relações semânticas dos termos.

A etapa onde é utilizado o algoritmo de Similaridade de Cosseno com base nas pontuações encontradas retorna resultados bem mais precisos quanto a similaridade dos documentos extraídos.

Os dois métodos de Mineração de Dados compõem uma abordagem híbrida que procura complementar a deficiência que cada um apresenta e se apresenta como uma metodologia inovadora para a recomendação social.

O uso da Recomendação por *Tags* busca desta forma solucionar problemas típicos dos Sistemas de Recomendação abordados no Capítulo 6, onde a tentativa de modelar o comportamento e interesses do usuário em ambiente de recomendação apresenta-se como principal barreira acarretando em problemas como: Problema do Novo Usuário, Problema do Novo Item, Esparsialidade e Superespecialização. Os termos e palavras-chaves co-ocorrendo em documentos diferenciados, mas com contextos semelhantes aproximam da identificação daquilo que o usuário está procurando.

A Recomendação por *Tags* infere que a ocorrência freqüente de *tags* em um documento aponta para uma similaridade semântica, entretanto, este aspecto pode apresentar suas limitações seja por uma imprecisão lingüística ou excesso de ambigüidade textual. A utilização da abordagem híbrida de filtragem de informações por meio do TF-IDF e do algoritmo Similaridade de Cosseno contribuem na diminuição da redundância, ou eliminação da ambigüidade dos termos (detecção de sinonímia e detecção de polissemia). Uma técnica procurar complementar a deficiência da outra.

A contribuição maior deste trabalho é sem dúvida a criação de uma metodologia nova para ser adotada no processo de recomendação de Redes Sociais Baseadas em Localização e a construção de uma aplicação embarcada para a LBSN. Este ambiente *Web Social* ainda carece de pesquisas mais aprofundadas a consulta feita pelo usuário e o uso contínuo de técnicas de Mineração de Dados em Texto com resultados consolidados.

8.2 Resultados Alcançados

Este Trabalho de Dissertação contemplou uma abordagem geral do temas: Mineração de Dados e Sistemas de Recomendação; além de apresentar um estudo inicial sobre as Redes Sociais como um todo, pois o aspecto principal da pesquisa não se detém ao estudo aprofundado das Redes Sociais e sim a o efeito da recomendação social, em especial no ambiente da LBSN.

Para as Redes Sociais Baseadas em Localização alcançaram-se estudos a respeito dos trabalhos mais relevantes (ainda que poucos na área), principalmente aqueles em que existem as técnicas de Mineração de Dados envolvida, a caracterização da LBSN ou que se delimitem a uma pesquisa semelhante a este Trabalho de Dissertação.

Outras metas foram atingidas, tais como:

- Complementação de trabalhos já existentes na área;
- Criação de uma Abordagem Híbrida de Filtragem de Informação por meio dos métodos de Mineração de Dados;
- A resolução para problemas clássicos dos Sistemas de Recomendação;
- Criação de uma metodologia de recomendação social ainda nova para ser aplicada em LBSN, de forma que possa solucionar algumas deficiências;
- Aplicação e testes controlados da metodologia proposta em uma Rede Social Baseada em Localização;
- Publicação de artigo sobre o tema da referida dissertação em Conferência Internacional de Sistemas Inteligentes: *International Conference on Intelligent Systems, Modelling and Simulation* - ISMS 2013, Bangkok, 29-31 de January 2013.

8.3 Trabalhos Futuros

Como proposta para trabalhos futuros pretende-se:

- Aprimorar a coleta de dados na Rede Social;
- Melhorar a metodologia desenvolvida, sobretudo os algoritmos utilizados na Mineração de Dados;
- Estudo de outros métodos de Mineração de Dados que possam vir a agregar melhorias na recomendação social;
- Aprofundamento nas pesquisas que possam solucionar outras deficiências dos Sistemas de Recomendação;
- Criar uma interface própria para a realização da Recuperação de Dados em uma Rede Social Baseada em Localização, tendo em vista que a aplicação é apenas embarcada para a própria rede social do *Foursquare*;
- Por fim, desenvolver uma aplicação que seja independente de uma Rede Social Baseada em Localização específica através de técnicas que envolvam práticas do paradigma da Inteligência Coletiva.

REFERÊNCIAS

ABBASSI, Z., MIRROKNI, V.S. **A recommender system based on local random walks and spectral methods.** In Advances in Web Mining and Web Usage Analysis – 9th International Workshop on Knowledge Discovery on the Web, WebKDD 2007 and 1st International Workshop on Social Networks Analysis, SNA-KDD 2007, San Jose, CA, USA, August 12-15, 2007, Revised Papers (LNAI 5439), pages 139–153, 2009.

ADOMAVICIUS, Gediminas. Tuzhilin, A. **Toward next generation of recommender systems: A survey of the state-ofthe-art and possible extensions**. IEEE Transactions on ledge and data engineering, vol. 17, no. 6, june 2005.

AHA, David W., KIBLER, D. and Albert, Marc K. **Instance-Based Learning Algorithms**. Machine Learning, 6, pages 37-66, 1991.

AMO, S. **Curso de Data Mining, Programa de Mestrado em Ciência da Computação**, Universidade Federal de Uberlândia, 2003. Disponível em: http://www.deamo.prof.ufu.br/CursoDM.html.

ANGELONI, M. T. (Org). **Organizações do conhecimento: infra-estrutura, pessoas e tecnologias**. São Paulo: Saraiva, 2008.

ANSARI, A. et al. **Internet Recommendation Systems. Journal of Marketing Research**, v.37, n.3, p. 363-375, Aug. 2000.

ARGYRIOU, L., PATRIKAKIS, C. Z., PORTER, S. CM., PAPAOULAKIS, N., ANDROULAKI, Christina. **Using media related user profiles to personalize multimedia access over social networks**. Proceeding SBNMA '11 Proceedings of the 2011 ACM workshop on Social and behavioural networked media access Pages 9-14.

BAEZA-YATES, R. Ribeiro-Neto, B. **Modern Information Retrieval**. ACM Pres, 1998.

BAEZA-YATES, R., POBLETE, Barbara . **A website mining model centered on user queries.** In Web and Mining – Joint International Workshops, EWMF 2005 and KDO 2005, Porto, Portugal, October 3 and 7, 2005, Revised Selected Papers (LNAI 4289), pages 1–17, 2006.

BALABANOVIC, M., & SHOHAM, Y. (1997). **Fab: Content-based collaborative recommendation.** Communications of the ACM , 40(3), 88-89.

BAO, Jie., ZHENG, Yu., MOKBEL, Mohamed F. **Location-based and preference-aware recommendation using sparse geo-social networking data**. SIGSPATIAL/GIS 2012: 199-208.

BANK, J. and COLE, B. **Calculating the Jaccard Similarity Coeficient with Map Reduce for Entity Pairs in Wikipedia**. Disponível em: http://www.infosci.cornell.edu/weblab/papers/Bank2008.pdf.

BARTAL, Alon., SASSON, Elan., RAVID, Gilad. **Predicting Links in Social Networks Using Text Mining and SNA**. ASONAM 2009: 131-136.

BASTOS, Valéria M.. **"Ambiente de Descoberta de Conhecimento na Web para a Língua Portuguesa"**. Tese de Doutorado, COPPE/UFRJ, 2006.

BELKIN, Nicholas J, CROFT, W. Bruce. **Information Filtering and Information Retrieval: Two sides of the same coin**. Comm. ACM. vol. 35, nº 12, 1992.

BELVIN, N. J.; CROFT, W. B.(1992). **Information Filtering and Information Retrieval: two sides of the same coin?**. Communications of the ACM, New York, v.35, n.12, p. 29, Dec.

BENEVENUTO, F., Almeida, V. **Uma Análise Empírica de Interações em Redes Sociais**. In Proceedings of the XXIV Concurso de teses e dissertações (CTD). Natal, Brazil. July, 2011.

BENTLEY, J. **Multidimensional binary search trees used for associative searching**. Communications of the ACM, Vol.18, pages 509-517, 1975.

BERENDT., B., MOBASHER, B., NAKAGAWA, M., SPILIOPOULOU, M. **The impact of site structure and user environment on session reconstruction in web usage analysis**. In WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profiles, 4th International Workshop, Edmonton, Canada, July 23, 2002, Revised Papers (LNAI 2703), pages 159–179, 2003.

BEZERRA, B. L. D. e CARVALHO, F. A. T de. **Information Filtering based on Modal Symbolic Objects**. Information Processing Letters, aceito em 2003, a ser publicado em 2004.

BHAGAT, S., CORMODE, G., ROZENBAUM, I. **Applying link-based classigication to label blogs**. In Advances in Web Mining and Web Usage Analysis – 9th International Workshop on Knowledge Discovery on the Web, WebKDD 2007 and 1st International Workshop on Social Networks Analysis, SNA-KDD 2007, San Jose, CA, USA, August 12-15, 2007, Revised Papers (LNAI 5439), pages 97–117, 2009.

BHATIA, M.P.S.,GAUR, P. **Statistical approach for community mining in social networks.Service Operations and Logistics, and Informatics**, 2008. IEEE/SOLI 2008. IEEE International Conference on 12-15 Oct. 2008.Volume:1 p.: 207-211.

BOYD, D. Why Youth (Heart) **Social Network Sites: The Role of Networked Publics in Teenage Social Life**. Cambridge, MA, 2007.

BRAGA, L. P. V. **Introdução a mineração de dados**. 2ª ed. Editora e-Papers, 2005.

BRIN, S. & PAGE, L. "**The anatomy of a large scale Web Search Engine"**. In Seventh International World Wide Web Conference, Brisbane, Australia, 1998.

BURKE, R.: 2002**, 'Knowledge-based Recommender Systems'**. In: A. Kent (ed.): Encyclopedia of Library and Information Systems. Vol. 69, Supplement 32.

CARLONE, Domenico., ARROYO, Daniel Ortiz. **Semantically Oriented Sentiment Mining in Location-Based Social Network Spaces**. FQAS 2011: 234-245.

CARMEL, David., ROITMAN, Haggai., YOM-TOV, Elad. **Social Bookmark Weighting for Search and Recommendation**. VLDB JournalVolume 19 Issue 6, 2010.

CATTUTO et al. **Semantic Analysis of Tag Similarity measures in Collaborative Tagging Systems**. Proceedings of the 3rd Workshop on Ontology Learnin and Population OLP3 Patras, 2008.

CAZELLA, Sílvio César; REATEGUI, Eliseo Berni. **Sistemas de Recomendação**. São. Leopoldo: XXV Congresso da Sociedade Brasileira de Computação, 2005.

CAZELLA, Sílvio César., REATEGUI, Eliseo Berni., ALVARES, Luis Otávio Campos. **E-commerce recommenders' authority: applying the user's opinion relevance in recommender systems**. WebMedia 2006: 71-78.

CHAN, P. **A non-invasive learning approach to building web user profiles**. Workshop on Web usage analysis and user profiling, Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, August 1999, pages 7-12.

CHENG, Z., CAVERLEE, J., LEE, K., and SUI, D. (2011). **Exploring Millions of Footprints in Location Sharing Services**. In Proc. of ICWSM'11.

CHI, Ed H., ROSIEN, A., HEER, J. **Lumberjack: Intelligent discovery and analysis of web user traffic composition**. In WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profiles, 4th International Workshop, Edmonton, Canada, July 23, 2002, Revised Papers (LNAI 2703), pages 1–16, 2003.

CHO, E., MYERS, S., and LESKOVEC, J. **Friendship and Mobility: User Movement.** In Location-Based Social Networks. In Proc. of KDD'11, 2011.

CLAYPOOL, M., GOKHALE, A., MIRANDA, T., MURNIVOK, P., NETES, D. e SARTIN, M. **Combining Content-based and Collaborative Filters in an Online Newspaper**. In Proceedings of ACM SIGIR Workshop on Recommender Systems, August 19 1999.

COOLEY, R. MOBASHER, B., SRIVASTAVA, J. **"Web mining: information and pattern Discovery on the World Wide Web"**. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, 1997.

COSTA, Helen., BENEVENUTO, Fabricio and MERSCHMANN, Luiz. **Detecting Tip Spam in Location-based Social Networks**. In Proceedings of the ACM Symposium on Applied Computing (SAC'13). Coimbra, Portugal, March 2013.

COOLEY, R. MOBASHER, B., SRIVASTAVA, J. **Data preparation for mining world wide web browsing patterns**. Knowledge and Information Systems, 1:5–32, 1999.

COVER, T. M., e HART, P. E. (1967). **Nearest Neighbor Classifiers**. IEEE Transactions on Computers, 23-11, November, 1974, pages 1179-1184.

CRAMER, H., ROST, M., and HOLMQUIST L. E. **Performing a Check-in: Emerging Practices, Norms and 'Conflicts' in Location-Sharing Using**. 2011

DAVENPORT, Thomas H.; PRUSAK, Laurence. **Conhecimento Empresarial – como as organizações gerencial o seu capital intelectual**. Rio de Janeiro: Elsevier, 2003.

DENNING, P. J. **Eletronic Junk**. Communications of the ACM, New York, v.25, n.3, p. 163-165, Mar. 1982.

FAYYAD, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996c) **From data mining to knowledge discovery: an overview**. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, Advances in Knowledge Discovery and Data Mining, pages 1-34. AAAI Press / MIT Press, Menlo Park, CA.

FELDMAN-BIANCO, Bela (org.). **Antropologia das sociedades contemporâneas: métodos**. São Paulo: Global, 1987.

FELDMAN, R. & SANGER, J. **The Text Mining Handbook-Advanced Approaches in Analyzing Unstructured Data**, USA: New York, 2007.

FERRARI, Laura., ROSI, Alberto., MAMEI, Marco., ZAMBONELLI, Franco. **Extracting urban patterns from location-based social networks**. GIS-LBSN 2011: 9-16.

FREITAS, Juliana Gonçalves. **Uma ferramenta para clusterização de perfis de usuários baseada em dados qualitativos**. Centro de Ciências Exatas e Tecnológicas. Universidade do Vale do Rio Sinos, 2008.

GAMBS, Sébastien., HEEN, Olivier., POTIN, Christophe. **A comparative privacy analysis of geosocial networks**. SPRINGL 2011: 33-40.

GEYER-SCHULZ, A., HAHSLER, M. **Comparing two recommender algorithms with the help of recommendations by peers.** In WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profiles, 4th International Workshop, Edmonton, Canada, July 23, 2002, Revised Papers (LNAI 2703), pages 137–158, 2003.

GOLDBERG, D., Nichols, D., Oki, B. M., Terry, D. 1992. **Using collaborative filtering to weave an information tapestry**. Communications of the ACM, 35(12), 61-70.

GOECKS, J. and SHAVLIK, J. W. (2000). **Learning users' interests by unobtrusively observing their normal behavior**. In Proceedings of the ACM Intelligent User Interfaces Conference (IUI), Jan. 2000, pages 129-132.

GOLDSCHIMIDT, Ronaldo; Passos, Emmanuel. **Data Mining: um guia prático**. Rio de Janeiro: Elsevier, 2005 – 4ª edição.

HAIR-JR., J. F. et al (2005). **Análise multivariada de dados**. Capítulo 9 - Análise de Agrupamentos. pp. 381-419. Bookman.

HAMASAKI, M., and Takeda, H. **Find better friends? re-configuration of personal networks by neighborhood matchmaker method**. In SWAFT (2003), pp. 73–76.

HAN, J. **"OLAP Mining: An integration of OLAP with Data Mining"**. School of Computing Science, Simon Fraser University, British Columbia, Canada, 2000.

HAYES, C., AVESANI, P., BOJARS, U. **Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection**. In From Web to Social Web: Discovering and Deploying User and Content Profiles – Workshop on Web Mining, WebMine 2006, Berlin, Germany, September 18, 2006, Revised Selected and Invited Papers (LNAI 4737), pages 1–20, 2007.

HEARST, M.A. **Untangling Text Data Mining**. In the Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.

HERLOCKER, J. L. et al. **Evaluating Collaborative Filtering Recommender Systems.** ACM Transactions on Information Systems, New York, v.22, n.1, p. 5-53. Jan. 2004.

HEYMANN, P., KOUTRIKA, G., and GARCIA-MOLINA, H. 2008. **Can Social Bookmarking Improve Web Search?**. In WSDM '08: Proceedings of the intl. conf. on Web search and web data mining.

JAMALI, M., ABOLHASSANI, H. **Different aspects of social network analysis**. In: Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence (WI'06) (Washington, DC, USA), IEEE Computer Society, pp 66–72.

JIN, Lei., LONG, Xuelian X., JOSHI, James B.D. **Towards understanding Residential Privacy by Analyzing User' Activities in Foursquare**. In 2012 Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS'12), Held in conjunction with CCS 2012, Raleigh, NC, USA.

KIEFER, P., STEIN, K., SCHLIEDER, C. **Visibility analysis on the web using co-visibilities and semantic networks**. In Semantics, Web and Mining – Joint International Workshops, EWMF 2005 and KDO 2005, Porto, Portugal, October 3 and 7, 2005, Revised Selected Papers (LNAI 4289), pages 34–50, 2006.

KIM, Hyoung Rae., CHAN, Philip K. **Personalized search results with user interest hierarchies learnt from bookmarks**. In Advances in Web Mining and Web Usage Analysis – 7th International Workshop on Knowledge Discovery on theWeb, WebKDD 2005, Chicago, IL, USA, August 21, 2005, Revised Papers (LNAI 4198), pages 158–176, 2006.

KIM, Heung-Nam., ROCZNIAK, Andrew., LÉVY, Pierre., EL-SADDIK, Abdulmotaleb. **Social media filtering based on collaborative tagging in semantic space**. Multimedia Tools Appl. 56(1): 63-89 (2012).

KING, Irwin., LYU, Michael R., MA, Hao. **Introduction to social recommendation**. WWW 2010: 1355-1356.

KLEINBERG, J.M. **"Authoritative Sources in a Hyper-linked Enviroment"**. In Proc. ACM-SIAM Symposyum on Discrete Alghoritms, 1998.

KOSALA, R., BLOCKEEL, H. **"Web mining research: a survey"**. SIG KDD Explorations, vol.2, pp. 1-15, 2000.

KOSCH, H. **"Distributed Multimedia Database Technologies supported by MPEG-7 and MPEG-21,"** CRC Press. 280 pages. November 2003. ISBN: 0-849-31854-8.

KUMAR, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. **"The Web and Social Networks"**. IEEE Computer, vol.35, no.11, 2002, pp.32-36.

LABIDI, S. **Managing Multi-Expertise in the Design of Cooperative Knowledge-Based Systems**. Proceedings of the IEEE Knowledge and Data Exchange Workshop (KDEX' 97). New Port Beach, Lose Angeles, USA. November, 1997.

LAMPE, Cliff., ELLISON, Nicole B., STEINFIELD, Charles. **A familiar face(book): profile elements as signals in an online social network**. CHI 2007: 435-444.

LI, Nan., CHEN, Guanling. **Analysis of a Location-Based Social Network**. CSE (4) 2009: 263-270.

LICCARDI, I., Ounnas, A., Pau, R., Massey, E., Kinnunen, P., Lewthwaite, S., Midy, M., and Sarkar, C. 2007. **The role of social networks in students' learning experiences**. SIGCSE Bull. 39, 4 (Dec. 2007), 224-237.

LOPS, Pasquale., GEMMIS, Marco de., SEMERARO, Giovanni., FEDELUCIO, Narducci., MUSTO, Cataldo. **Leveraging the linkedin social network data for extracting content-based user profiles**. RecSys 2011: 293-296.

MACHADO, Aydano P. **Mineração de texto em redes sociais aplicada à educação Distância**. Colabora – Revista digital da CVA – Ricesu, vol. 6, n. 23, jul, 2010.

MASSA, P. et. al. **Using trust in Recommender Systems: an experimental analysis**. In: INTERNATIONAL CONFERENCE ON TRUST, iTrust, 2004. Proceedings... [S.1.], n.24,p. 253-276, 2005.

MELIA-SEGUI, J., ZHANG, R., BART, E., PRICE, B., BRDICZKA, O. **Activity duration analysis for context-aware services using foursquare check-ins**. Proceedings of the 2012 International Workshop on Self-aware Internet of Things, co-located with the 9th ACM International Conference on Autonomic Computing (ICAC 2012), Pages 13 - 18. San Jose, California (USA), September 2012.

MONTANER, M.; López, B.; LA ROSA, J. L. **A Taxonomy of Recommender Agents on the Internet.** Artificial Intelligence Review.

MORITA, M., & Shinoda, Y. (1994). **Information filtering based on user behavior analysis and best match text retrieval**. In Proceedings of SIGIR, Dublin, Ireland, ACM Press, 1994, pages 272-281.

MOTOYAMA, M. and VARGHESE, G., **I seek you: searching and matching individuals in social networks"**. in Proceedings of the eleventh international workshop on Web information and data management, ser. WIDM '09, 2009.

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. **Criação de conhecimento na empresa: comoas empresas japonesas geram a dinâmica da inovação**. 14 ed. Rio de Janeiro:Campus, 1997.

NOULAS, A., SCELLATO, S., MASCOLO, C., and PONTIL, M. (2011). **Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks**. In Proc. of SMW'11.

OARD, D. and KIM, J. **Implicit Feedback for Recommender Systems**. In AAAI Technical Report WS-98-08: Workshop on Recommender Systems, July 27, Madison, WI.

OKA, Mizuki., MATSUO, Yutaka. **Mining Scholarly Semantic Networks from the Web**. IV 2008: 349-355.

PAL, Sankar K., TALWAR, V., MITRA, P. **"Web Mining in Soft Computing Framework: Relevant, State of the Art and Future Directions"**, 2000.

RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTROM, P., & RIEDL, J. T. (1994). **GroupLens: An open architecture for collaborative filtering of Netnews**. Proceedings of the Conference on Computer Supported Cooperative Work. ACM. xi+464, 175-86.

RESNICK, P. and VARIAN, H. R. 1997. **Recommender systems**. Communications of the ACM, 40(3), 56-58.

RIEDL, J. et al. **Combining Collaborative Filtering with Personal Agent for Better Recommendations**. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 16.; INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE CONFERENCE, 11., 1999. Proceedings... Menl Park, CA: American Association for Artificial Intelligence [S.1.], 1999. p. 439-446.

ROBERTSON, R. Sparck, J. **Relevance weighting of search terms**. Journal of the American Society for Information Science, 27:129--146, 1976.

ROMSAIYUD, W., PREMCHAISWADI, W., **"Applying mining fuzzy sequential patterns technique to predict the leadership in social networks"**. ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2011 9th International Conference on, On page(s): 134 – 137.

SALTON, G.; BUCKLEY, C. **Term-Weighting Approaches in Automatic Text Retrieval**. Information Processing and Management, v. 24, n. 5, 1988. p.513-523.

SANTOS, André Luís Silva dos. **Um modelo de sistema de filtragem híbrida para um ambiente colaborativo de ensino aprendizagem**. Dissertação de Mestrado, Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2008.

SARWAR, B. M. et al. **Analysis of Recommendation Algorithms for e-commerce**. In: ACM CONFERENCE ON E-COMMERCE, 2., 2000, Minneapolis. Proceedings... New York: ACM Press, 2000. p. 158-167.

SCHAFER, J.B., KONSTAN, J.A., and RIEDL, J. (1999). **Recommender Systems in E-Commerce.** In Proceedings of the First ACM Conference on Electronic Commerce (pp. 158–166). Denver, CO: ACM Press.

SCHAFER, J.B; KONSTAN, J.; Riedl, J. **E-Commerce recommendation applications. Data Mining and Knowledge Discovery.** [S.1], v5,n. 1-2, p. 115-153, Jan. 2001.

SCHENKEL, R., Crecelius, T., KACIMI, M., Michel, S., NEUMANN, T., PARREIRA, J. X., WEIKUM, G. (2008) **Efficient top-k querying over social-tagging networks**. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 523-530.

SFERRA, Heloisa Helena; CORREA, Ângela M. C. Jorge. **Conceitos e Aplicações de Data Mining.** Jul/Dez de 2003, Revista Ciência & Tecnologia, PP. 19-34.

SHARDANAND, U., & MAES, P. (1995). **Social information filtering: Algorithms for automating 'Word of Mouth'.** Proceedings of CHI '95. ACM. xx+598, 210-17.

SIGURBJÖRNSSON, Börkur., and VAN ZWOL, Roelof. **Flickr Tag Recommendation based on Collective Knowledge**. Proceedings of the 17th International World Wide Web Conference (WWW'08). 2008.

SINGH, Y., CHAUHAN, A. Singh. **Neural Networks In Data Mining.** Journal of Theoretical and Applied information Technology, 2005, p 37 to 42.

SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., Tan, P.N. “Web usage mining: Discovery and applications of usage patterns from Web data”. SIG KDD Explorations, 2000.

SULLIVAN, D. (2000). **The need for text mining in business intelligence**. DM Review, Dec. 2000. Disponível em: http://www.dmreview.com/master.cfm.

TAN, A.-H. **"Text mining: The state of the art and the challenges"**. In Proceddings, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, Beijing, 1999. pp.65-70.

TAN, S. **An effective refinement strategy for KNN text classifier**, Expert Systems with Applications 30, 2006, pp:290–298.

TEIXEIRA, I. R. **Um Método de Aprendizagem Ativa em Sistemas de Filtragem Colaborativa.** Tese de Mestrado em Inteligência Artificial, Universidade Federal de Pernambuco, 2002.

THURAISINGHAM, B. **Data mining: technologies, techniques, tools, and trends**. CRC Press, Boca Raton, Florida, 1999.

TSO-SUTTER, K.H.L., MARINHO, L.B., and SCHMIDT-THIEME, L.. **“Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms”,** Proceedings of the 2008 ACM symposium on Applied computing, ACM, USA, 2008, pp. 1995-1999.

UTARD, H., FÜRNKRANZ, J**. Link-local features for hypertext classification**. In Semantics, Web and Mining – Joint International Workshops, EWMF 2005 and KDO 2005, Porto, Portugal, October 3 and 7, 2005, Revised Selected Papers (LNAI 4289), pages 51–64, 2006.

VASCONCELOS, M., RICCI, S., Almeida, J, BENEVENUTO, F., and ALMEIDA, V. (2012). **Tips, Dones and Todos: Uncovering User Profiles in Foursquare**. In Proc. of the WSDM’12.

WASSERMAN, S. et al. **Social Network Analysis: Methods and Applications**. [S.l.]: Cambridge University Press, 1994.

WELLMAN, B. **Physical place and cyberplace: The rise of personalized networking**. International Journal of Urban and Regional Research 25, Special Issue on “Networks, Class and Place” (2001). Edited by Talja Blokland and Mike Savage.

WEISS, S. M. e INDURKHYA, N. **"Predictive Data Mining"**. California, USA: Morgan Kaufmann, 1998.

YANG, L., RAHI, A. **Dynamic clustering of web search results**. In Computational Science and Its Applications – ICCSA 2003, International Conference, Montreal, Canada, May 18-21, 2003, Proceedings, Part I (LNCS 2667), pages 153–159, 2003.

YE, Mao., YING, Peifeng, LEE, Wang-Chien., LEE, Dik Lun. **"Exploiting Geographical Influence for Collaborative Point-of-Interest Recommendation"**, In Proceedings of the ACM International Conference on Research & Development on Information Retrieval (SIGIR'11), pages: 325-344, 2011.

ZAIANE, O. R., Han, J., LI, Z. –N., CHEE, S.H., & CHIANG, J. **“Multimidia data miner: a system prototype for multimedia data miner”.** In Proc. ACM SIGMOD Intl. Conf. on Management of Data, pages 581-583, 1998”.

ZHAO, S., DU, N., NAUERZ, A., ZHANG, X., YUAN, Q., and R. FU. **Improved recommendation based on collaborative tagging behaviors**. In Proceedings of the 13th international conference on Intelligent user interfaces, 2008.

ZHANG, J., and Ackerman, M. S. **Searching for expertise in social networks: a simulation of potential strategies**. In GROUP (2005), ACM, pp. 71–80.

APÊNDICES

APÊNDICE A – Corpo principal da classe em *Python* Autenticacao

```
import unittest
import urllib
from foursquare import OAuthHandler, API, BasicAuthHandler,
FoursquareError
from models import Tip, User


class TestAuthentication(unittest.TestCase):
    CLIENT_ID = 'YOUR_CLIENT_ID'
    CLIENT_SECRET = 'YOUR_CLIENT_SECRET'
    REDIRECT_URI = 'YOUR_CALLBACK'

    def _test_create_OAuthHandler(self):
        auth = OAuthHandler(TestAuthentication.CLIENT_ID,
                            TestAuthentication.CLIENT_SECRET,
                           TestAuthentication.REDIRECT_URI)
        self.assertEquals(auth._client_id,
TestAuthentication.CLIENT_ID)
        self.assertEquals(auth._client_secret,
TestAuthentication.CLIENT_SECRET)
        self.assertEquals(auth.callback,
TestAuthentication.REDIRECT_URI)

    def _test_get_authorization_url(self):
        auth = OAuthHandler(TestAuthentication.CLIENT_ID,
TestAuthentication.CLIENT_SECRET,
           TestAuthentication.REDIRECT_URI)
        self.assertEquals(auth.get_authorization_url(),

('https://foursquare.com/oauth2/authenticate?redirect_uri=%s' +
            '&response_type=code&client_id=%s')
            % (urllib.quote(self.REDIRECT_URI).replace('/', '%2F'),
self.CLIENT_ID)
        )

    def _test_get_access_token(self):
        auth = OAuthHandler(TestAuthentication.CLIENT_ID,
TestAuthentication.CLIENT_SECRET,
            TestAuthentication.REDIRECT_URI)
        code = 'YOUR_CODE'
        self.assert_(auth.get_access_token(code) is not None)
```

APÊNDICE B – Corpo principal da classe em *Python* Extractor

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
import os
import json
import urllib2

class Extractor():

url = "https://api.foursquare.com/v2/venues/search?ll=40.7,-
74&near=Maranh%C3%A3o&query=caranguejo&oauth_token=334VCSXMHXJI2BIW1TK
RXOBHM33PWFGEC1DGHFFPLK244IES&v=20130108"

json_string = urllib2.urlopen(url).read()
data = json.loads(json_string)
i = 0

l = []

for i in range (3):
      coleta = data['response']['venues'][i]['name']
      l.append(coleta)

                                    .
                                    .
                                    .
dicionario = dict(a34=b34,a35=b35,a36=b36)
print dicionario

x = open('dicionario_venues.txt','w')
x.write(json.dumps(dicionario))
x.close

_table = {"\u00e0" : "a", "\u00e1" : "a", "\u00e2" : "a","\u00e3" :
"a" , "\u00e4" : "a", "\u00c1" : "A", "\u00c0" : "A",
"\u00c2" : "A", "\u00c3" : "A",  "\u00c4" : "A", "\u00e9" : "e",
"\u00e8" : "e", "\u00ea" : "e", "\u00ea" : "e", "\u00c9" : "E",
"\u00c8" : "E", "\u00ca" : "E", "\u00cb" : "E", "\u00ed" : "i",
"\u00ec" : "i", "\u00ee" : "i", "\u00ef" : "i", "\u00cd" : "I",
"\u00cc" : "I", "\u00ce" : "I", "\u00cf" : "I", "\u00f3" : "o",
"\u00f2" : "o", "\u00f4" : "o", "\u00f5" : "o", "\u00f6" : "o",
"\u00d3" : "O", "\u00d2" : "O", "\u00d4" : "O", "\u00d5" : "O",
"\u00d6" : "O", "\u00fa" : "u","\u00f9" : "u", "\u00fb" : "u",
"\u00fc" : "u", "\u00da" : "U", "\u00d9" : "U", "\u00db" : "U",
"\u00e7" : "c", "\u00c7" : "C", "\u00f1" : "n", "\u00d1" : "N",
"\u0026" : "&", "\u0027" : "'" }


def asciize(s):
    """
    Converts a entire string to a ASCII only string.

    string
        The string to be converted.
    """
    for original, plain in _table.items():
        s = s.replace(original, plain)
    return s
```

APÊNDICE C – Corpo principal da classe em *Python* Calculo_TF_IDF

```
# -*- coding: utf-8 -*-

import sys
import os
import json
from math import log

class TF_IDF():
QUERY_TERMS = sys.argv[1:]

def tf(term, doc, normalize=True):
    doc = doc.lower().split()
    if normalize:
        return doc.count(term.lower()) / float(len(doc))
    else:
        return doc.count(term.lower()) / 1.0


def idf(term, corpus):
    num_texts_with_term = len([True for text in corpus if
term.lower()in text.lower().split()])

    # tf-idf calc involves multiplying against a tf value less than 0,
so it's important
    # to return a value greater than 1 for consistent scoring.
(Multiplying two values
    # less than 1 returns a value less than each of them)

    try:
        return 1.0 + log(float(len(corpus)) / num_texts_with_term)
    except ZeroDivisionError:
        return 1.0


def tf_idf(term, doc, corpus):
    return tf(term, doc) * idf(term, corpus)

# Score queries by calculating cumulative tf_idf score for each term
in query

query_scores = dict.fromkeys(corpus,0)
l = []
for term in [t.lower() for t in QUERY_TERMS]:
    for doc in sorted(corpus):
        print 'TF(%s): %s' % (doc, term), tf(term, corpus[doc])
    print 'IDF: %s' % (term, ), idf(term, corpus.values())
    print

    for doc in sorted(corpus):
        score = tf_idf(term, corpus[doc], corpus)
        print 'TF-IDF(%s): %s' % (doc, term), score
        query_scores[doc] += score
        l.append ((doc, score))

    print
    d = dict (l)
    x = open('pontos_tf_idf.txt','a')
    x.write(json.dumps(d))
    x.close
```

APÊNDICE D – Corpo principal da classe em *Python* Calculo_Sim_Cosseno

```
import nltk
import os
import json

class Sim_Cosseno():

for doc in sorted(corpus):
    distance = nltk.cluster.util.cosine_distance(v1,v2)
    print distance

    distance = dict (l)
    x = open('pontos_sim_cosseno.txt','a')
    x.write(json.dumps(d))
    x.close
```

APÊNDICE E – Corpo principal da classe em *Python* Ranking_Tag

```
from numpy import *

class ranking_tag():

def knn(k,data,dataClass,inputs):

      nInputs = shape(inputs)[0]
      closest = zeros(nInputs)

      for n in range(nInputs):
            # Compute distances
            distances = sum((data-inputs[n,:])**2,axis=1)

            # Identify the nearest neighbours
            indices = argsort(distances,axis=0)

            classes = unique(dataClass[indices[:k]])
            if len(classes)==1:
                  closest[n] = unique(classes)
            else:
                  counts = zeros(max(classes)+1)
                  for i in range(k):
                        counts[dataClass[indices[i]]] += 1
                  closest[n] = max(counts)

      return closest
```

ANEXO

ANEXO 1 – Parte do Corpus de documento coletado para implementação

"a1": "Casquinha de Caranguejo", "a3": "Fabio", "a2": "Edson", "a5": "Recanto Baiano", "a4": "Novo Chico","a20": "Melhor caranguejo e cerveja beeem gelada...", "a15": "Caranguejo no leite de coco delicioso! O arroz de toucinho e o melhor da cidade.",
"a22": "O melhor caranguejo atendimento perfeito", "a23": "O pastel de carne com geleia de pimenta, a patola de caranguejo a vinagrete, a salada, o file do sol, affff tudo aqui e perfeito! Atendimento excelente e cerveja geladissima. Top!!!", "a24": "Melhor caranguejo do Brasil! Aproveite bem.",
"a25": "File de pescada ao molho de camarao: Supimpa!Casquinha de caranguejo tambem e uma boa pedida.", "a26": "Evite a torta de caranguejo congelada", "a27": "Otima opcao para quem gosta de caranguejo.", "a28": "Caranguejo, pexe frito e farofa.. Delicia!!",
"a29": "Linguica de camarao e excelente e a coxinha de caranguejo e 10.",
"a19": "O arroz de toucinho com caranguejo e de comer rezando =)",
"a17": "Caranguejo bom, mas a fila, nem um pouco... se prepare para as senhas!", "a18": "Melhor tora de caranguejo da", "a21": "O melhor peixe frito e caranguejo", "a16": "Experimente a camaroada, caranguejada ao leite de coco.. Com certeza a melhor da cidade!! A entrada de patinhas de caranguejo tb e muito boa!!",
"a7": "A coxinha de caranguejo do Ferreiro Grill e matadora. Faz o maior sucesso em Aracaju. Sera que no de Sao Luis segue o mesmo padrao?", "a6": "Caranguejo toc-toc...maravilhoso!!!", "a9": "Caranguejo,camarao e matei meu desejo de ostra com limao, local sob a agua dos lencois maranhenses voce almoca e se perde com a vista do mar", "a11": "A patinha de caranguejo e uma delicia",
"a31": "Risoto de frutos do mar e patinha de caranguejo ao vinagrete. Hummmmm dilicia...", "a10": "O Miele tem o melhor caranguejo de Sao Luis e o melhor casquinho. Peixes grelhados. Pescada amarela de primeira qualidade.", "a33": "Caranguejo muuuuito bom!", "a14": "Caranguejo TOC TOC... melhor tira gosto da casa!", "a13": "O caranguejo ao molho e um pecado e com esse esse arroz de tocinho entao...",
"a12": "Melhor casquinha de caranguejo da ilha, com feijoada e pagode aos sabados.", "a8": "Camaroada, arroz de cuxa, torta de caranguejo... Otimo restaurante maranhense...", "a30": "A casquinha de caranguejo e a melhor!", "a32": "Local onde vende os maiores e melhores caranguejos!","a36": "Quiosque de Caranguejo", "a35": "Caranguejo Bar", "a34": "Casa Do Carangueijo","e1": "Camaroada, arroz de cuxa, torta de caranguejo... Otimo restaurante maranhense...","e24": "Restaurante Do Jack", "e25": "Restaurante Tudo de Bom", "e22": "Restaurante Maracangalha", "e29": "Restaurante O Comilao", "e9": "Antigamente Restaurante & Bar", "e8": "Restaurante dos Arcos", "e5": "Restaurante Escola Senac", "e4": "Restaurante Universitario - UFMA", "e7": "Restaurante SESC Deodoro", "e6": "Restaurante do Franca", "e3": "Porto Seguro - Bar & Restaurante", "e2": "Restaurante Gula-Gula", "e19": "Restaurante Tio Patinhas", "e18": "Restaurante", "e26": "Restaurante Alencar", "e27": "Restaurante Malagueta", "e20": "Restaurante Tempero de Mae", "e21": "Restaurante Ponto Certo", "e31": "Crioula's Bar E Restaurante", "e30": "Restaurante - Porto Norte", "e11": "Restaurante Paladar", "e10": "Restaurante Pequim", "e13": "Restaurante Caiu do Ceu TV DIFUSORA", "e12": "Restaurante Cabana do Sol", "e15": "Restaurante Comida Caseira", "e14": "Restaurante Dona Nina", "e17": "Restaurante D' Antiga mente", "e16": "Restaurante Cabana do Sol", "e23": "Restaurante Antigamente (CEMAR)", "e28": "Restaurante Baiao dos Dois","c2": "Evite a torta de caranguejo