Tiago Novaes Angelo

Extrator de Conhecimento Coletivo: uma ferramenta para Democracia Participativa

Campinas
2014

Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Tiago Novaes Angelo

Extrator de Conhecimento Coletivo: uma ferramenta para Democracia Participativa

Orientador: Prof. Dr. Ricardo Ribeiro Gudwin
Coorientador: Prof. Dr. Cesar José Bonjuani Pagan

Dissertação de mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas para a obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: Automação.

Este exemplar corresponde à versão final da dissertação de mestrado defendida pelo aluno Tiago Novaes Angelo e orientada pelo Prof. Dr. Ricardo Ribeiro Gudwin e Prof. Dr. Cesar José Bonjuani Pagan

---

Campinas
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

An43e

Angelo, Tiago Novaes, 1983-
Extrator de conhecimento coletivo : uma ferramenta para democracia participativa / Tiago Novaes Angelo. – Campinas, SP : [s.n.], 2014.

Orientador: Ricardo Ribeiro Gudwin.
Coorientador: Cesar José Bonjuani Pagan.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Comunicações digitais. 2. Redes de informação - Aspectos sociais. 3. Gestão participativa. 4. Processamento de linguagem natural (Computação). 5. Redes complexas. I. Gudwin, Ricardo Ribeiro,1967-. II. Pagan, Cesar José Bonjuani,1962-. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Extractor Collective Knowledge : a tool for participatory democracy
Palavras-chave em inglês:
Digital communications
Information networks - Social aspects
Participatory management
Natural language processing (Computer)
Complex networks
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Ricardo Ribeiro Gudwin [Orientador]
João Luís Garcia Rosa
Guilherme Palermo Coelho
Data de defesa: 29-08-2014
Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Tiago Novaes Angelo

Data da Defesa: 29 de agosto de 2014

Título da Tese: "Extrator de Conhecimento Coletivo: Uma Ferramenta para Democracia Participativa"

Prof. Dr. Ricardo Ribeiro Gudwin (Presidente): _____

Prof. Dr. João Luís Garcia Rosa: _____

Prof. Dr. Guilherme Palermo Coelho: _____

Resumo

O surgimento das Tecnologias de Comunicação e Informação trouxe uma nova perspectiva para o fortalecimento da democracia nas sociedades modernas. A democracia representativa, modelo predominante nas sociedades atuais, atravessa uma crise de credibilidade cuja principal consequência é o afastamento do cidadão na participação política, enfraquecendo os ideais democráticos. Neste contexto, a tecnologia surge como possibilidade para construção de um novo modelo de participação popular que resgate uma cidadania mais ativa, inaugurando o que se denomina de democracia digital. O objetivo desta pesquisa foi desenvolver e implementar uma ferramenta, denominada "Extrator de Conhecimento Coletivo", com o propósito de conhecer o que um coletivo pensa a respeito de sua realidade a partir de pequenos relatos de seus participantes, dando voz à população num processo de democracia participativa. Os fundamentos teóricos baseiam-se em métodos de mineração de dados, sumarizadores extrativos e redes complexas. A ferramenta foi implementada e testada usando um banco de dados formado por opiniões de clientes a respeito de suas estadas em um Hotel. Os resultados trazem evidências significativas que o algoritmo pode satisfazer a seus propósitos. Para trabalhos futuros, a proposta é que o Extrator de Conhecimento Coletivo seja o núcleo de processamento de dados de um espaço virtual onde a população possa se expressar e exercer ativamente sua cidadania.

Palavras-chave: Democracia Digital. Democracia Participativa. Democracia. Processamento de Linguagem Natural. Redes Complexas. Sumarizadores extrativos. Mineração de dados.

Abstract

The emergence of Information and Communication Technologies brought a new perspective to the strengthening of democracy in modern societies. The representative democracy, prevalent model in today's societies, crosses a crisis of credibility whose main consequence is the removal of citizen participation in politics, weakening democratic ideals. In this context, technology emerges as a possibility for construction of a new model of popular participation to rescue more active citizenship, inaugurating what is called digital democracy. The objective of this research was to develop and implement a tool called "Collective Knowledge Extractor", with the purpose of knowing what the collective thinks about his reality through small reports of its participants, giving voice to the people in a process of participatory democracy. The theoretical foundations are based on methods of data mining, extractive summarizers and complex networks. The tool was implemented and tested using a database consisting of customer reviews about their stay in a Hotel. The results provide significant evidences that the algorithm can satisfy their purposes. For future work, the proposal is that the Collective Knowledge Extractor might be the core data processing of a virtual space where people can express themselves and actively exercise their citizenship.

Keywords: Digital Democracy. Participatory Democracy. Democracy. Natural Language Processing. Complex Networks. Extractive summarizers. Data mining.

Sumário

Aos meus pais, Sônia e Abelardo, responsáveis por me ensinar sobre amor, ética e honestidade e que fizeram tantos sacrifícios pela minha formação acadêmica, acreditando no meu potencial e me apoiando e orientando incondicionalmente em todas as minhas decisões.

Agradecimentos

Agradeço,

A Deus, por iluminar meu caminho ao mostrar que o sentido da vida está no amor ao próximo.

Ao Professor Cesar Pagan, exemplo de dedicação e ética profissional, responsável por ter me despertado o interesse em transformar o mundo em um lugar melhor, pela brilhante orientação deste trabalho, pela amizade e apoio incondicional, pela ética e honestidade e pelo papel de Mestre, Professor e Educador que me marcou profundamente e me guiará profissionalmente.

Ao Professor Ricardo Gudwin, pelo apoio, orientação e dedicação dispensados em todos os momentos.

Ao Professor Romis Attux pelo papel de educador, pelas sugestões e contribuições que expandiram o meu conhecimento e pela amizade construída ao longo desses anos.

Ao Professor Léo Pini, pela excelente orientação docente e profissional ao longo do meu mestrado.

Aos professores da FEEC-Unicamp por todo o conhecimento e oportunidade que me ofereceram.

Aos professores da PUC-Campinas que me ensinaram que o bem-estar do ser humano deve ser prioridade na minha atuação profissional.

Aos meus amigos, em especial aos Totós, que souberam entender minha ausência nos momentos em que estive mais atarefado e me apoiaram com a mais nobre amizade.

Aos membros da banca examinadora pelos comentários, sugestões e contribuições.

Ao CNPQ, pela concessão da bolsa de estudo.

A todos que de alguma forma contribuíram para esse trabalho e pela minha formação profissional e pessoal.

Do rio que tudo arrasta se diz que é violento. Mas ninguém diz violentas as margens que o comprimem.

Bertolt Brecht

Lista de Figuras

Lista de Tabelas

Introdução

O desenvolvimento científico-tecnológico tem acarretado rápidas e profundas transformações na sociedade contemporânea. A ampliação do conhecimento através do desenvolvimento da ciência propiciou, e ainda propicia, a criação de uma série de instrumentos e produtos tecnológicos cuja função é mediar dialeticamente a relação entre indivíduo e sociedade levando tanto a transformações econômicas, sociais, culturais e políticas, como também transformações na subjetividade de cada agente social definindo novas formas de agir, pensar e comunicar (HALL, 2006). Tal condição implica em uma reflexão crítica sobre os caminhos do saber científico, principalmente no que diz respeito à criação de aparatos tecnológicos, já que podem diretamente impactar em benefícios ou malefícios para o progresso social e individual.

Dentre as principais inovações tecnológicas que transformam a sociedade contemporânea pode-se destacar a invenção do *transistor* que permitiu o aparecimento de dispositivos como microprocessadores e computadores digitais e marcou o início do que é denominado atualmente "Era da Informação" (NARAYANAMURTI; ODUMOSU; VINSEL, 2013). Estes sistemas levaram a grandes avanços nas áreas de telecomunicações como a criação da fibra óptica e redes de comunicação interligadas globalmente impactando em profundas mudanças principalmente no que diz respeito ao dinamismo das comunicações e ao fluxo de informações. Assim, ao longo das últimas décadas, observou-se um forte avanço na capacidade de criar, manipular e armazenar a informação em sistemas eletrônicos e digitais ocasionando o desenvolvimento de áreas como a Robótica, Inteligência Artificial, Processamento de Linguagem Natural, entre outras.

Desta forma, o desenvolvimento tecnológico alterou rapidamente os paradigmas sociais ao longo das últimas décadas. O conhecimento passou a ser a principal fonte de produtividade na economia contemporânea, a qual é fortemente dependente da capacidade de gerar, processar e aplicar eficientemente a informação; a riqueza de uma nação passou a ser medida pelo seu acesso à tecnologia e capacidade de desenvolvimento na área; elementos sociais, culturais e políticos também passaram a depender fortemente da informação e do suporte tecnológico para propagá-la (SILVA; CORREIA; LIMA, 2010).

No entanto, diante do desenvolvimento de sistemas cada vez mais robustos e potentes, capazes de alterar profundamente a sociedade e os indivíduos nela inseridos, faz-se necessário refletir sobre para que e para quem toda esta tecnologia está sendo direcionada. Feenberg (1991) aponta que dentro de uma conjuntura capitalista, a tecnologia tende a ser vista como uma produção neutra, a-histórica, desvinculada da instância social e voltada unicamente para racionalidade

técnica. A consequência desta visão, de que a tecnologia está desvinculada do seu contexto social e histórico, é que esta se torna um potente instrumento de dominação e concentração de poder indo no sentido inverso do seu uso para o benefício de todos os indivíduos da sociedade. Porém, ao se tomar consciência de que a tecnologia é uma construção social, a trajetória da inovação científica pode ser redirecionada e ir ao encontro, dependendo do interesse daquele que a produz, da emancipação humana. A tecnologia torna-se, neste sentido, uma "promessa de liberdade" (FEENBERG, 1991).

Dentro dessa lógica, é possível concluir que na atual era da informação, dominar o conhecimento científico-tecnológico e ser capaz de criar instrumentos e máquinas, significa estar na vanguarda da transformação social ou simplesmente da reprodução do *status quo*. Por este motivo, a produção científica, em especial aquela voltada para as ciências aplicadas, não deve estar deslocada do seu contexto histórico e social. A Engenharia é um poderoso instrumento de mudança social já que está no cerne das tecnologias da informação; porém, seu progresso nem sempre coincide com a evolução do bem-estar. Por isso, ao se produzir ciência o engenheiro deve estar consciente de que a ciência é uma criação humana, logo, histórica e social e também de que toda produção científica impacta, positiva ou negativamente, a sociedade.

Por outro lado, para que a ciência gere benefícios e desenvolvimento social, o conhecimento deve ser construído dentro do contexto em que será utilizado. A produção de conhecimento, e não apenas a cópia do que é produzido em outras realidades, é um fator determinante de autonomia de um país e de desenvolvimento social já que coloca como prioridade os problemas reais de uma população, diminuindo as severas consequências da dependência tecnológica (CARVALHO, 1997).

Desta forma, refletindo sobre a responsabilidade em se produzir ciência na área da Engenharia, as atuais demandas da população e o impacto que a tecnologia pode ter na sociedade brasileira, a proposta deste estudo é trazer a tecnologia para mais próximo da gestão pública, desenvolvendo um sistema que seja capaz de beneficiar a democracia brasileira, carente de participação direta, e possa levar sem intermédios ou representantes os anseios de uma comunidade aos seus dirigentes políticos.

Aproximar a tecnologia à gestão pública é uma demanda cada vez mais constante no mundo contemporâneo. Em tempos de globalização, descentralização e inovação tecnológica, os governantes são cada vez mais pressionados a apresentar respostas aos desafios de uma sociedade cujo mundo real está sendo amplificado pela tecnologia através da virtualização e do dinamismo das comunicações. Porém, possuir tecnologia e informação não é garantia de transformação social capaz de gerar mudanças na gestão pública, a qual demanda uma efetiva participação democrática dos atores envolvidos. A informação deve ser sempre relevante e acessível, respondendo aos anseios populares e garantindo que a comunicação promova encontros entre interlocutores diferentes, de forma que seja possível mobilizá-los para ações efetivas verdadeiramente democráticas (RIBEIRO; SOPHIA; GRIGÓRIO, 2007).

Historicamente, o uso da tecnologia e da informática na gestão pública brasileira se concentra principalmente no oferecimento de programas para automatizar os serviços públicos e também na área de finanças públicas. O tema é de tamanha relevância que, desde 1995, ocorre o Congresso de Informática Pública (CONIP), um fórum brasileiro de debates sobre o uso da tecnologia no

setor público (DINIZ, 2005). Porém, analisando os temas discutidos neste congresso desde sua primeira edição, pouco é falado a respeito da promoção da democracia usando tecnologia.

No atual momento histórico pelo qual o Brasil atravessa, pensar na questão da democracia como instrumento de participação ativa e direta do cidadão é uma demanda cada vez mais crescente. Em meados de Junho de 2013, um movimento social denominado MPL, Movimento do Passe Livre, promoveu uma série de manifestações na cidade de São Paulo com o objetivo de revogar o aumento da passagem do transporte público (MARADEI, 2013). Tais manifestações foram duramente reprimidas pela polícia militar, ganhando, assim, destaques nacional e internacional e sendo o estopim para uma onda de manifestações em várias cidades brasileiras. O movimento chamou a atenção não só pelo caráter reivindicatório, mas também pela rapidez com que as pessoas foram mobilizadas, devido, principalmente, a facilidade que as redes sociais, como o Facebook, promoveram para divulgação das informações.

Neste contexto, a discussão sobre o uso da tecnologia como instrumento de promoção da democracia faz-se extremamente pertinente. Alguns pontos podem ser levantados:

- Será que os Movimentos Sociais de fato representam na íntegra as reinvindicações de todas as pessoas que se manifestaram nas ruas?

- As medidas não seriam mais eficazes e as respostas viriam mais rápidas caso os governantes soubessem diretamente as reinvindicações de todos os manifestantes?

- Num mundo onde a virtualização está se tornando regra e a capacidade de processamento da informação é cada vez maior, não seria possível desenvolver um instrumento que "ouvisse" cada um desses manifestantes e fosse capaz de dizer o que eles de fato estão querendo?

Refletindo a respeito dessas indagações, seria muito interessante se existisse um espaço virtual capaz de agregar discussões, compilar ideias e mostrar aquelas que emergem nesse universo, dando possibilidades não só para expressão dos cidadãos como também abrindo espaço para uma participação ativa sobre a tomada de decisão política, num processo democrático sem mediadores, direto e participativo.

É imerso nesta demanda sócio-político brasileira e fundamentado no rápido desenvolvimento científico-tecnológico que está propiciando a criação de sistemas inteligentes e robustos para processar dados, que este estudo tem como pretensão lançar sementes para a criação de um ambiente virtual que possibilite a promoção de uma democracia mais direta e participativa a partir do desenvolvimento de um instrumento que seja capaz de compilar um conjunto muito grande de textos e retirar as ideias que prevalecem e emergem.

Duas condições foram essenciais para concretização deste estudo: a primeira foi ter clareza sobre os conceitos de democracia, como ela é vista nos dias atuais e quais são as influências da tecnologia para sua promoção. Em segundo buscou-se na literatura científica respaldos para desenvolver uma tecnologia de processamento de dados que trabalhasse com linguagem natural e engenharia do conhecimento. Assim, unindo conhecimentos da área de redes complexas, Processamento de Linguagem Natural, mineração de dados e Inteligência Artificial, além de outras afins como a linguística e a psicologia, foi proposto o desenvolvimento de um software denominado "Extrator de Conhecimento Coletivo" (ECC) cujo objetivo é buscar o conhecimento emergente dado um conjunto muito grande de pequenos discursos sobre um determinado tema ou realidade.

Para abordar todas estas questões, o presente trabalho está organizado em seis capítulos: no primeiro capítulo é realizada uma revisão sobre os conceitos de democracia, desde a democracia clássica até a democracia digital, destacando a criação de métodos e instrumentos tecnológicos que visam a promoção de modelos democráticos; no segundo capítulo é feita uma revisão sobre redes complexas, uma teoria que se utiliza da matemática de grafos para modelar sistemas complexos reais; no terceiro capítulo é feita uma revisão sobre algumas técnicas de sumarização extrativa, método de mineração de dados e extração de frases que inspiraram a arquitetura do ECC; o quarto capítulo apresenta a arquitetura proposta para o ECC, bem como detalhes de sua implementação; o quinto capítulo expõe uma aplicação prática do ECC, apresentando os resultados do processamento de um banco de dados formado por opiniões de clientes de um hotel; por fim, o sexto capítulo apresenta as conclusões e as perspectivas futuras.

Capítulo 1

Democracia

O surgimento das novas Tecnologias de Informação e Comunicação (TIC) vem transformando significativamente o panorama das modernas sociedades democráticas principalmente em relação à participação popular no exercício da cidadania. Antigos anseios da democracia clássica abandonados ao longo dos anos como a participação direta dos cidadãos nos negócios públicos vêm sendo resgatados graças ao desenvolvimento destas tecnologias, inaugurando um novo marco da democracia: a democracia digital.

Por ser um conceito relativamente novo, a vasta quantidade de experiências em democracia digital que surgiram nos últimos anos contemplam um amplo espectro de aplicações empíricas ainda pouco teorizadas e carentes de definições formais, o que torna seu estudo um desafio uma vez que são pouco conhecidos os impactos a médio e longo prazo de tais iniciativas na promoção de uma sociedade de fato mais democrática. Neste sentido, aproximar tais aplicações dos estudos acadêmicos é uma etapa necessária para seu amadurecimento como uma ciência que gera um conhecimento em prol da democratização efetiva da sociedade. Assim, algumas questões devem ser cuidadosamente observadas ao se pensar no desenvolvimento de ferramentas em democracia digital, dentre elas compreender o que é democracia, suas diferentes visões e conceitos, e qual o seu papel ao longo da história.

O presente capítulo iniciará fazendo um resumo sobre o que é democracia e quais suas principais vertentes filosóficas, resgatando sua história e analisando seu papel como alicerce de um modelo de sociedade. Em seguida, apresentará quais as limitações dos modelos democráticos atuais e as dificuldades em torná-los efetivos. Por fim, apresentará como o desenvolvimento das tecnologias de informação e comunicação podem ajudar na superação destas dificuldades, citando algumas experiências que são o estado da arte em democracia digital.

1.1 Origens da Democracia

Definir exatamente quando a democracia surgiu não é uma tarefa simples. Os primeiros vestígios de práticas democráticas datam de muito antes da Grécia Antiga em Sociedades do Oriente que se organizavam em assembleias e elegiam representantes numa espécie de "democracia primitiva" que parecia antecipar as assembleias populares gregas. Heródoto, no terceiro livro

de suas Histórias, ao descrever um debate entre conspiradores persas, relata que a democracia não somente já era conhecida como havia sido inventada pelos persas, inimigos históricos dos gregos na Antiguidade (CANFORA, 2008).

Porém, foi na Grécia Antiga durante o século V a.C. que a democracia passou a fazer parte do pensamento político e filosófico, principalmente na cidade-Estado de Atenas onde se estabeleceu o primeiro governo democrático conhecido, liderado por Péricles. Neste período, toda vida política ateniense acontecia nas Assembleias Populares e, apesar de ser um regime popular, nomeá-lo de "democracia" foi uma ação dos opositores ao regime. "Cracia", oriunda da palavra *kratos* significa, em seu sentido literal, força violenta, portanto, dizer que era um regime democrático era uma crítica ao que os opositores consideravam um governo popular de caráter violento. Além disso, "demos" não era uma representação de toda população ateniense. Pelo contrário, era um termo que se referia as pessoas consideradas "sem posses", a maioria da população livre que não fazia parte nem da oligarquia e nem dos militares e, ainda assim, eram a minoria da população grega, a qual 3/4 era formada por escravos que não gozavam de nenhum direito político enquanto apenas $\frac{1}{4}$ era constituída por homens livres. A democracia começou de fato a nascer quando a cidadania (direito de participação política), antes direito apenas dos oligarcas e militares, foi estendida também aos "sem posses", incluindo todos os homens livres da sociedade ateniense (CANFORA, 2008).

O regime democrático ateniense não só foi alvo de críticas de opositores políticos como também dos principais pensadores da época como Platão e Aristóteles, os quais buscavam uma razão filosófica sobre as formas ideais de governança. Aristoteles (1986 (c.320 BC) apud CUNNINGHAM, 2002) foi quem mais se aprofundou nesta busca e encabeçou uma pesquisa cujo objetivo era descrever e esboçar as histórias de todos os sistemas políticos conhecidos na época. Deste trabalho, concluiu que havia seis possíveis formas de governo: realeza, tirania, aristocracia, oligarquia, politéia e a democracia; e que, independente da forma de governo, o princípio da maioria prevalece, ou seja, sempre será a maioria dominante que terá em mãos o poder, seja a maioria dos ricos (na oligarquia) ou a maioria do povo (na democracia). Além disso, Aristóteles concluiu que a realeza seria a forma ideal de governo e considerava a democracia uma deturpação do modo de governar, pois com ela não seria possível governar em vista do bem comum. Porém, de todas as possíveis deturpações políticas, a democracia era o desvio de governo mais "tolerável" já que era possível obter benefícios através das experiências coletivas e seria mais fácil governar quando a maioria absoluta não está descontente (CUNNINGHAM, 2002).

Apesar de nascer já sendo alvo de críticas, o modelo democrático ateniense manteve-se e se aprimorou durante a Grécia Antiga, mostrando seu caráter coletivo e de respeito às leis e a justiça. Diversos outros pensadores contribuíram para o desenvolvimento da democracia ateniense, cujo legado denominou-se anos mais tarde de democracia clássica e influenciou o surgimento de modelos políticos favoráveis e contrários a este (CUNNINGHAM, 2002). Porém, alguns fundamentos da democracia ateniense, como a participação direta e ativa da população nas decisões políticas foram sendo deixados de lado em prol de modelos mais representativos. Atualmente, com o advento da tecnologia, surge a possibilidade de rediscutir estes fundamentos e resgatar a cidadania mais ativa e direta. Vale então, neste ponto, uma breve explanação sobre

as características e práticas da democracia clássica, assunto do próximo tópico.

1.2 A Democracia Ateniense

Na cidade-estado de Atenas, a vida do cidadão estava estritamente ligada à vida da *polis*. O princípio que prevalecia era o da "virtude cívica", o qual estabelecia que todos os cidadãos deveriam dedicar-se à cidade subordinando a vida privada às questões públicas e ao bem comum, gerando uma distinção entre cidadão e indivíduo onde os diretos coletivos eram superiores aos direitos individuais. Desta forma, grande parte da vida dos atenienses era dedicada às questões públicas, caracterizada por uma cidadania ativa e de um processo de auto-governo cujo princípio de governança era a participação cidadã direta. Na prática, o cidadão ateniense dedicava-se a encontros para debater e decidir as leis, os quais se constituíam em discussões livres e irrestritas onde todos tinham direitos iguais para falar em uma assembleia soberana. Após estes debates, decisões eram tomadas a partir do poder de convencimento dos argumentos e as leis decididas tornavam-se leis do estado (HELD, 2006).

O uso das assembleias soberanas foi fortalecido no governo de Péricles (V a.C) e passaram a constituir o coração do sistema democrático onde os cidadãos se reuniam para discutir os problemas da *polis* e criar as leis. Dentre estas assembleias, a principal era a Ekklésia, ou Assembleia do Povo, local onde todos os cidadãos tinham direto a palavra e ao voto e deliberavam questões ligadas a defesa do país, escolha de magistrados para cargos políticos, julgavam questões de traição e direitos políticos entre outras. Outra assembleia era dedicada aos "suplícios", onde todos os cidadãos poderiam colocar assuntos de caráter público ou privado para discussão ou deliberação. Para que as sessões fossem válidas, as assembleias exigiam quórum mínimo de 6 mil cidadãos e as votações eram decididas de forma direta, por maioria simples, com cada votante levantando ou não a mão para o alto (MENEZES, 2010).

Segundo Held (2006) a participação direta dos cidadãos foi a principal característica da política ateniense e tornou-se o fundamento básico da democracia clássica perdurando até a queda de Atenas quando surgiram os impérios, os "estados fortes" e o poderio militar. Apesar do aparente sumiço, os ideais democráticos foram profundamente difundidos na época do Império Romano e voltaram à tona no período iluminista quando diversos pensadores resgataram os escritos gregos sobre democracia e descreveram uma vasta gama de modelos democráticos, uns como crítica à democracia direta ateniense, outros como tentativa de superação das suas dificuldades e outros ainda como resgate à participação popular direta.

Neste resgate iluminista, uma das modificações mais significativas do modelo de democracia clássica e que obteve grande sucesso no decorrer da história moderna foi a transferência da participação direta dos cidadãos para um sistema centralizado de representação política. Enquanto na Grécia ateniense a esfera pública era o espaço de tomada de decisão e deliberações políticas cujos participantes eram os cidadãos, nos Estados Modernos a participação política passou a ser mediada por um corpo independente de políticos profissionais legitimados pelos cidadãos, gerando uma cisão que colocou de um lado a esfera pública, cujo objetivo é escolher os representantes e, de outro lado, a esfera política a qual de fato define e delibera os rumos da sociedade. A este modelo, predominante nas democracias atuais, denomina-se Democracia

Representativa.

1.3 Democracia Participativa *versus* democracia Representativa

A democracia nos Estados Modernos é oriunda de um misto do desenvolvimento da ideia de representação com a de igualdade de direto e realiza-se na forma de um governo representativo. John Stuart Mill teria sido o primeiro pensador a associar a ideia de governo representativo com democracia ao vislumbrar a inviabilidade técnica da participação direta de todos no governo, nascendo assim o modelo de democracia representativa (CASTANHO, 2012).

De acordo com Mill (1861 apud HELD, 2006), a ideia grega de *polis* não é sustentável na sociedade moderna. A noção de auto-governo seria insensata para qualquer comunidade que exceda o tamanho de uma cidade pequena já que nem todos conseguiriam participar de todas as discussões. O trabalho de organizar e coordenar multidões seria muito complexo para que todos participem diretamente das decisões e, há, inclusive, limitações físicas já que não é possível um lugar onde todos os cidadãos de uma grande cidade ou do próprio país se reúnam para discutir e deliberar leis. Ainda, segundo Mill, em um regime governado por todos os cidadãos há um perigo constante das pessoas mais sábias e capazes serem silenciadas pela falta de conhecimento, habilidade e experiência das maiorias. Desta forma, conclui que as dificuldades da participação direta em governos democráticos só podem ser contornadas pela escolha de governantes através de eleição popular, em um regime denominado Democracia Representativa.

Nesta forma de governo, o poder político passa do próprio cidadão para a figura de um representante, estabelecendo um vínculo representante-representado que pode se expressar politicamente de três formas: Primeiro como uma técnica onde a vontade de uma pessoa é atribuída a outra pessoa ou à coletividade; a segunda como um instrumento para exprimir a vontade do representado onde o representante é uma espécie de porta-voz do representado; e a terceira como uma substituição, onde o representante substitui o povo, atribuindo a ele sua vontade (CASTANHO, 2012).

Segundo Young (2006), nas sociedades modernas a existência da representação é uma condição necessária pois muitas vezes a ação política do cidadão está vinculada a processos que ocorrem em vários e diferentes locais e instituições não sendo possível estar presente em todos os organismos deliberativos e debater todas as ideias que irão afetar suas vidas. Além disso, ainda que nem sempre tenha suas expectativas atendidas, o cidadão espera que outros pensem em situações como a dele e o represente em fóruns ou órgãos deliberativos.

A democracia representativa permitiu que os fundamentos democráticos se mantivessem em uma sociedade que se tornava cada vez mais numerosa e complexa, porém, o papel da representação política nem sempre se estabeleceu de forma que os reais interesses da população fossem de fato atendidos, o que gerou inúmeras críticas a este modelo. Alguns autores atribuem a atual crise do sistema democrático aos modelos representativos e suas instituições como partidos políticos (BENNETT; ENTMAN, 2001) (BUCY; GREGSON, 2001) (MIGUEL, 2006) (PHARR; PUTNAM, 2000) cuja consequência é o descrédito político e o enfraquecimento da cidadania.

Barber (2003), defensor da democracia participativa e crítico ao modelo de democracia representativa, argumenta que a representação é incompatível com a liberdade, igualdade e justiça social pois:

"aliena a vontade política em detrimento do genuíno auto-governo, (...) prejudica a capacidade da comunidade de atuar como um instrumento regulador, (...) e impede a formação de um público participativo"

Neste sentido, defende um modelo que denomina de "democracia forte", um sistema democrático semelhante às assembleias gregas onde o cidadão atue de forma direta nas decisões. Já para Gomes (2006), o modelo representativo está levando a uma crise do sistema democrático pois ao separar a esfera política da esfera pública esvazia o interesse dos cidadãos na participação política e faz com que a política contemporânea não seja capaz de satisfazer os requisitos da democracia em seu sentido mais próprio. Dentre as consequências desta cisão está um sentimento de desinteresse no cidadão que não percebe os efeitos de suas ações visto a forma que o Estado age em relação aos seus anseios. Tal sentimento reforça-se pela percepção de que a indústria da notícia, do lobby e da consultoria política têm muito mais eficácia sobre a esfera política do que os próprios cidadãos. Neste sentido há uma marginalização do papel do cidadão gerando uma sensação de ineficácia da ação política que contribui para arruinar as condições da participação cívica. Além destes fatores, o desinteresse também é reforçado pela péssima imagem pública da sociedade política, percebida como voltada unicamente ao jogo de interesses próprios, de grupos/partidos ou a interesses não-públicos puramente econômicos.

Segundo Coleman e Blumler (2009), apesar de vivermos numa época de grandes possibilidades e oportunidades de mudar governos, as pessoas nunca se sentiram tão frustradas e desapontadas com a falta de capacidade de fazer qualquer diferença na política e nas decisões públicas. Observou em entrevistas, grupos focados e pesquisas que as pessoas repetitivamente queixam-se da sensação de estarem excluídas, não serem ouvidas e serem desrespeitadas, tornando-se meros espectadores do processo político o qual rapidamente torna-se pouco confiável. Lavalle, Houtzager e Castello (2006) observam alguns indícios que também podem indicar uma crise do modelo democrático representativo, dentre os quais destacam-se a volatilidade do eleitorado, a queda nos patamares de comparecimento nas urnas e um descrédito generalizado nas instituições públicas.

Em relação ao modelo de democracia representativa adotado no Brasil, os problemas se voltam à atuação dos partidos políticos, excessiva burocratização e distanciamento dos interesses sociais (PERISSINOTTO; FUKS, 2002). Segundo Menezes (2010), a grande quantidade de partidos políticos, os quais deveriam gerar uma pluralidade de ideias e princípios, acabou estabelecendo uma situação na qual partidos nanicos se tornaram "legendas de aluguel" constituindo negócios privados que não vão ao encontro da representatividade dos interesses públicos. Tal fato favoreceu o surgimento de grupos de pressão que atuam nos bastidores sem qualquer regulamentação ou responsabilidade a partir de tráfico de influência. Desta forma, pouco ou nada se representa da vontade pública e a política se torna um mero jogo de interesses onde seus cidadãos que, sem poder de atuar, mudar sua realidade ou serem ouvidos, passam a desacreditar

no sistema político acarretando duas consequências graves: a apatia política e a corrupção. A apatia é uma consequência da falta de confiança nos políticos e nas suas capacidades de defender os interesses da coletividade uma vez que há uma percepção de que suas ações ficam confinadas aos interesses de seus partidos e às estratégias pessoais as quais asseguram sua eleição. Já a corrupção, percebida como a falta de compromisso público dos políticos, que também não acreditam e por isso não valorizam a democracia, é uma consequência também da própria apatia política uma vez que esta acarreta na diminuição da fiscalização e regulamentação dos processos políticos por parte dos cidadãos.

Outra crítica ao modelo representativo está exatamente no processo de representação. Segundo Pitkin (1971), para que as formas de representatividade política sejam respeitadas, o representante deve ser capaz de deliberar sempre a favor de um coletivo atendendo a uma vasta gama de interesses. Porém, neste processo, toda decisão que tomar estará influenciada por seus valores morais já que ele próprio também é um representado. Isto exigiria um alto grau de neutralidade e um excelente canal de comunicação entre as esferas pública e política, porém o natural distanciamento destas na democracia representativa prejudica o processo de representação.

Apesar da hipótese de que o modelo representativo esteja levando a uma crise dos sistemas democráticos, opor a democracia clássica como um sistema ideal de participação popular e a democracia representativa como fonte de interesses individuais e tensão entre governantes e governados não é um caminho frutífero para o fortalecimento e estabelecimento da democracia. Cada modelo foi pensado em um contexto histórico diferente cujos desafios sociais eram muitas vezes de outra natureza e reproduzi-los nem sempre é a melhor solução para os desafios atuais. Além disso, os modelos não são práticas puras, eles se misturam em diversas situações: por exemplo, na Grécia ateniense havia certo grau de representação quando magistrados eram eleitos para tarefas executivas ou mesmo nas democracias representativas modernas a participação popular direta pode ser vista em práticas como plebiscitos e referendos (MENEZES, 2010). Desta forma, refletir sobre as vantagens e desvantagens de cada modelo contextualizando sua prática em um período sócio-histórico, atento às demandas sociais e às ferramentas tecnológicas, é um possível caminho para tentativas de superação da crise do sistema democrático.

E desta reflexão é que surgiram os modelos de democracia participativa, tal como o modelo de "democracia forte" de Barber (2003), já citado anteriormente, o qual se apresenta como um sistema de democracia participava que une conceito da democracia clássica e representativa cuja prática só é possível graças às novas tecnologias de informação e comunicação.

Os modelos de democracia participativa são relativamente novos e passaram a ganhar força a partir do inicio deste século graças ao surgimento de ferramentas tecnológicas que permitiram aproximar pessoas e dar espaço para uma participação popular mais direta nas decisões políticas. Na democracia participativa, garantir a atuação cidadã é o objetivo principal e, para isto, duas condições são necessárias: a acessibilidade do poder público à participação popular e um eficiente processo informativo já que o cidadão deve ser capaz de formular suas próprias questões e não apenas responder a questões prontas como, por exemplo, num processo de plebiscito ou referendo que acabam por isolar o cidadão da possibilidade de atuar diretamente nas decisões políticas. A votação, nesta visão, é um tipo muito pobre de participação política direta pelo simples fato

de deliberar questões pré-fabricadas. Num contexto de democracia participativa, deseja-se o surgimento de uma opinião coletiva a partir de debates e discussões em um ambiente rico em informações. Para isto, o uso das tecnologias de informação e comunicação é um caminho lógico a seguir (DIJK, 2000).

Enquanto que na história os ideais de democracia direta foram sendo vencidos e tidos como inadequados a medida que a sociedade se tornava uma sociedade de massas e o Estado uma instituição altamente complexa, o desenvolvimento tecnológico trouxe à tona a possibilidade de reverter esse processo. O surgimento de uma nova infraestrutura de comunicação e processamento de dados como por exemplo a internet está trazendo novas esperanças para o estabelecimento de modelos alternativos de democracia que resgatem o verdadeiro significado de cidadania e devolvam ao cidadão o poder de decisão política (GOMES, 2005). E é neste contexto que, principalmente após o início da década passada, surgem diversas iniciativas de base tecnológica cujo objetivo é resgatar a participação cidadã ativa. A este conceito que une as TIC e os modelos de democracia denominou-se "democracia digital", assunto do próximo tópico.

1.4 A Democracia Digital

Apesar de ser um conceito que ganhou força com o advento tecnológico dos últimos anos, a democracia digital não é um tema novo na literatura científica. Segundo Vedel (2006), a história da democracia digital pode ser dividida em três "eras". A primeira começou por volta de 1950 com o surgimento das tecnologias de computação e automação as quais geraram uma promessa de criação de uma "máquina de governar" que processaria uma grande quantidade de dados e tomaria decisões racionais a respeito da administração pública. No entanto, na década de 60 essa ideia foi abandonada devido a críticas de que ela estaria muito mais direcionada à tecnocracia do que de fato a um governo eletrônico pois reduzia a política unicamente à sua prática. A segunda era começa entre o fim da década de 70 e inicio da década de 80 com o aparecimento da TV a cabo e dos computadores pessoais. Neste período, a televisão começou a ser usada como espaço para debates, discussões e interatividade com os cidadãos. Porém, a promoção de democracia nestes espaços ainda era muito primitiva devido a limitações próprias da tecnologia, restringindo-se a ambientes mais informativos e a pouca possibilidade de atuação por parte do cidadão. Já a terceira era é a que melhor pode se associar com o conceito atual de democracia digital. Ela se inicia em meados da década de 90 com a emergência das mídias digitais como a internet e as possibilidades da comunicação em rede fazendo surgir o ativismo cibernético com a criação de espaços de debate virtual, fóruns, blogs, entre outras formas de expressão democrática. Com a popularização destas mídias e o desenvolvimento de ferramentas de processamento de dados, a democracia digital passa de um conceito para uma práxis com reais possibilidades de transformar a atuação democrática.

Nos últimos anos o entusiasmo sobre as possibilidades que as mídias digitais vêm trazendo para o avanço do sistema democrático é generalizado, abrangendo os interesses desde os governos oficiais até os libertários antigovernamentais, o que reflete o grande potencial da democracia digital em promover as mais diferentes visões de democracia. Dahlberg (2011), em um estudo sobre as potencialidades da democracia digital que envolveu a análise de diversas aplicações

em diferentes países, concluiu que as iniciativas podem ser classificadas em quatro esboços teóricos: democracia digital liberal-individualista, democracia digital deliberativa, democracia digital contra-público e democracia digital marxista-autonomista. Não será escopo deste estudo definir estes quatro esboços, porém vale ressaltar que ao desenvolver tecnologia nesta área há necessidade de se ter clareza sobre qual sistema democrático deseja-se promover uma vez que estudos como o de Dahlberg mostram que a democracia digital está carregada de conceitos ideológicos que terão impacto sobre a própria sociedade.

Mesmo que haja um grande esforço na criação de aplicações, teorizar a democracia digital ainda não é uma tarefa perto de ser concluída. Embora desde meados do século passado a tecnologia venha sendo pensada como suporte dos ideais democráticos, ainda não há um consenso sobre a definição deste conceito visto que as aplicações muitas vezes são de preocupação muito mais pragmática e os modelos democráticos são diversos filosoficamente e ideologicamente. Nem mesmo o nome democracia digital é unanimidade no meio acadêmico. Há autores que se referem ao mesmo fenômeno como e-democracia (LIDÉN, 2013), democracia eletrônica (SPIRAKIS; SPIRAKI; NIKOLOPOULOS, 2010), ciberdemocracia (GRONLUND; HORAN, 2005) (LÉVY, 2002), e-gov (ROMAN; MILLER, 2013), entre outras. No entanto, este estudo focará na definição de dois dos mais influentes autores em democracia digital: Steven Cliff, pesquisador pioneiro na área de democracia digital e criador do portal E-Democracy.org e o filósofo francês da cultura virtual contemporânea Pierre Lévy.

Em seu artigo *"E-democracy, E-Governance and Public Net-Work"*, Clift (2003) antes de definir o conceito, alerta que e-democracia, ou democracia eletrônica, não é votação direta via internet e muito menos campanha publicitária, mas sim:

"E-democracia é o uso de tecnologias de informação e comunicação e estratégias pelos setores democráticos dentro dos processos políticos de comunidade locais, estados/regiões, nações e mesmo num estágio global."

Os setores democráticos incluem os seguintes atores: Governos, representantes eleitos, mídia, partidos políticos e grupos de interesse, sociedade civil, organizações governamentais internacionais e os próprios cidadãos/votantes.

Já Lévy (2002) utiliza-se do conceito de ciberdemocracia para expressar o uso das tecnologias de informação e comunicação na promoção da democracia. Segundo o autor, ciberdemocracia é uma expressão que engloba o conceito de ciberespaço, local onde ocorre a comunicação e flui a informação, e o desejo de formas de governo genuinamente democráticas. O surgimento das tecnologias de rede que permitem a interligação mundial e ampliam a liberdade de comunicação estão estabelecendo um novo espaço público que virá a redefinir radicalmente as condições de governança. Para o autor este é o momento histórico mais propício para se repensar a atuação democrática já que a tecnologia poderá ser usada a favor de uma sociedade mais justa e igualitária. Dentre as promessas, Lévy (2002) destaca que graças às novas tecnologias:

"A própria natureza da cidadania democrática passa por uma profunda evolução, uma vez que caminha no sentido de um aprofundamento da liberdade: desenvolvimento do ciberativismo

à escala mundial, organização das cidades e regiões em comunidades inteligentes, em ágoras virtuais, governos eletrônicos cada vez mais transparentes ao serviço dos cidadãos e votos eletrônicos."

A ciberdemocracia seria não apenas uma forma de se pensar a democracia na Era da Informação, mas também uma forma de buscar o aperfeiçoamento social através do impulso à "inteligência coletiva". Uma vez que a noção de democracia remete aos direitos e liberdade do cidadão, cujas ideias devem ser debatidas e deliberadas em prol de algo comum a todos, a busca de uma regra ou ideia comum que expresse genuinamente o pensamento de uma coletividade é o que se denomina inteligência coletiva. A democracia, assim, implica numa inteligência coletiva a qual pode ser favorecida pelo uso das tecnologias de comunicação.

Apesar das promessas que a tecnologia tráz, muitos dos meios de comunicação atualmente pouco ajudam os povos a pensar coletivamente e criar soluções para seus problemas. No sistema político atual a informatização está servindo basicamente para simplificar os processos de burocratização e raramente está buscando formas criativas e inovadoras de tratar a informação de forma descentralizada, flexível, interativa e coletiva. Desta forma, é necessário conscientizar-se da necessidade de se explorar o enorme potencial transformador por parte das tecnologias (LÉVY, 1999).

A ciberdemocracia não seria uma forma de reforçar ou aprimorar o modelo democrático representativo, mas sim de incentivar e propiciar uma maior participação popular na vida da cidade explorando da melhor forma possível as ferramentas de comunicação contemporâneas, dentre elas, as tecnologias de rede e informação que compõe o ciberespaço, tornando-o o lugar de uma nova forma de democracia direta em grande escala. Assim, um uso socialmente mais rico da informática propiciaria aos grupos humanos os meios de reunir suas forças mentais e construir coletivos inteligentes que levem a uma democracia em tempo real (LÉVY, 1999).

Neste sentido, Lévy (1999) propõe a criação da "ágora virtual", uma hipótese utópica de uma plataforma virtual de democracia direta, a qual explora as potencialidades do ciberespaço na busca dos problemas, debates pluralistas, tomada de decisão coletiva e avaliação dos resultados sempre o mais próximo possível das comunidades envolvidas. Para que venha a se tornar realidade, há a necessidade do desenvolvimento de ferramentas de filtragem inteligente dos dados, navegação em meio a informação, simulação de sistemas complexos, comunicação transversal de tal forma que favoreça a tomada de decisão em coletivos heterogêneos e dispersos. A ágora virtual teria como objetivo facilitar a navegação no conhecimento permitindo a troca de saberes e a construção coletiva do sentido. Proporcionaria uma visão dinâmica das questões coletivas e a avaliação em tempo real de uma enorme quantidade de proposições, informações e processos em andamento. Além disso, como um instrumento de democracia participativa, o cidadão não mais seria um número que daria peso a um partido político ou a um representante, mas sim criaria diversidade e ampliaria o conhecimento contribuindo para o aperfeiçoamento da inteligência coletiva e resolução dos problemas comuns.

Apesar das particularidades de cada definição, tanto a ciberdemocracia quanto a e-democracia referem-se ao papel das novas tecnologias na promoção da democracia, ideia esta corroborada por Timonen (2013), o qual aponta cinco razões para usar as tecnologias e estratégias da demo-

cracia digital:

1. A democracia digital é essencialmente uma mídia social que permite aos governos e políticos se aproximarem dos cidadãos.

2. Permite que cidadãos com opiniões parecidas compartilhem ideias e organizem ações políticas.

3. A mídia digital fornece uma forma eficiente de disseminação de mensagens.

4. Redes digitais permitem o surgimento de muitas ideias e, portanto, dão poder aos cidadãos. Elas facilitam o acesso dos políticos a essas ideias de forma que possam torná-las políticas concretas.

5. Redes digitais permitem que os cidadãos se tornem verdadeiros responsáveis pelas decisões.

Indo além da definição, uma questão pertinente é avaliar o impacto das aplicações existentes nos processos democráticos. Com este objetivo, Silva (2005) desenvolveu um estudo envolvendo diversos projetos de democracia digital no Brasil e em outros países e concluiu que, baseado na capacidade do sistema em promover a participação popular nos negócios públicos, existem cinco possíveis graus de participação democrática virtual. Segue abaixo uma breve descrição destes graus:

- Primeiro grau de democracia digital: é caracterizado pela ênfase na disponibilidade de informação na prestação de serviços públicos. Nestes dispositivos, o governo busca suprir as necessidades de informação básica, serviços e bens públicos, enquanto o cidadão espera receber essas informações de forma rápida e sem transtornos. É um fluxo de informação unidirecional governo-cidadão e está intimamente relacionado com a melhoria de produtividade e otimização da máquina estatal.

- Segundo grau de democracia digital: aqui os dispositivos tecnológicos são criados para obter a opinião pública e utilizá-la para tomada de decisão política. A ideia é estabelecer um diálogo efetivo com a esfera pública, porém limitado apenas a um canal de sondagem de opinião sem garantias de que esta será de fato acatada.

- Terceiro grau de democracia digital: diz respeito às aplicações que atestam o princípio da transparência e da prestação de contas, permitindo o acesso da esfera pública aos dados do governo de forma que haja algum controle sobre as ações governamentais. Porém, apesar deste relativo controle, as decisões finais ainda serão tomadas unicamente pela esfera política.

- Quarto grau de democracia digital: são ferramentas capazes de propiciar discussões públicas que levarão a um consenso mútuo e implicarão em decisões concretas por parte da esfera política. São fundamentadas no diálogo aberto e livre dos participantes que devem propor reinvindicações e argumentos sobre problemas comuns. Dentro de um modelo de democracia representativa, este seria o grau mais elevado de participação democrática que o cidadão poderia participar.

- Quinto grau de democracia digital: este seria o grau mais intenso no que diz respeito à participação popular na tomada de decisão dos negócios públicos, aproximando-se dos modelos de democracia direta. Neste grau, a esfera pública e a esfera política se coincidiriam e a decisão pública passa a ter poder deliberativo. As ferramentas devem ser capazes de processar a informação pública e produzir decisão política, fazendo com que o cidadão de fato decida.

De acordo com Silva (2005), o desafio a ser enfrentado é melhorar o acesso dos cidadãos em cada um desses graus, o que envolve aspectos tecnológicos, sociais, econômicos e culturais. Para tal, há necessidade de gerar condições econômicas e políticas e desenvolver tecnologia, como a criação de sistemas, ferramentas, modelos, procedimentos e teorias que auxiliem atingir estes objetivos. No Brasil, os projetos existentes até o ano de 2005 estavam restritos ao primeiro e segundo graus, não havendo iniciativas que de fato atribuíssem um papel deliberativo às decisões populares. Já em outros países, algumas iniciativas se aproximam do terceiro e quarto graus de participação política. O próximo tópico tratará de apresentar o estado da arte das aplicações em democracia digital tanto no Brasil quanto ao redor do mundo.

1.5 Democracia Digital: o Estado da Arte

Muitas são as iniciativas em democracia digital ao redor do mundo cujo fator comum recai sobre as tentativas de aproximar as esferas pública e política. Em um estudo publicado em 2013 sobre o uso destas ferramentas, Timonen (2013) analisou como a democracia digital é empregada na União Europeia tanto por instituições governamentais como pelos próprios políticos. Dentre as aplicações, destaca a campanha de José Manuel Barroso para presidente da Comissão Europeia no ano de 2009 onde foi criado o TellBarroso (www.tellbarroso.eu), uma plataforma virtual de consulta pública onde os cidadãos podiam colocar propostas sobre ações que gostariam que fossem adotadas pela futura Comissão Europeia. A participação na consulta foi expressiva: mais de 150 mil pessoas participaram com 12 mil propostas, 500 mil visualizações e 130 milhões de impressões em mídias sociais. Timonen também descreve outros usos das mídias sociais em campanhas europeias como o caso do Partido Europeu Socialista que nas eleições para o Parlamento Europeu utilizou as mídias sociais para defender a criação de uma taxa de transação financeira, ganhando não só visibilidade como também conseguindo aprová-la posteriormente. Já em relação ao uso da tecnologia em sistemas de votação, cita o caso da Estônia que criou um sistema de votação eletrônica usando cartões inteligentes que permitem a votação via internet ou telefone celular.

O estudo de Timonen (2013) foi restrito a algumas aplicações recentes na União Européia e ressaltou as inúmeras possibilidades da democracia digital. Nchise (2012), publicou um estudo no qual analisou 158 artigos científicos que diziam respeito a democracia digital e traçou um panorama sobre o estado da arte, concluindo que:

- Os artigos em democracia digital tiveram um "boom" de publicações a partir do ano de 2003, saltando de 1 artigo publicado em 2001 para 23 em 2003.

- A maioria das aplicações encontram-se na Europa (64) e nas Américas (36). África, Austrália, Ásia e Oriente Médio juntos produziram 23 artigos. Os 35 artigos restantes não especificam a região.

- 59 artigos focam apenas em discussões teóricas sobre democracia digital. 84 relatam alguma aplicação prática e 15 apresentam opiniões e ponto de vista de acadêmicos sobre o tema.

- Sobre as aplicações, elas englobam 5 áreas: sistemas de voto eletrônico; sistemas de negociação eletrônica; sistemas de deliberação on-line; petições online e campanha virtual.

Em relação às aplicações no Brasil, os projetos em democracia digital vêm apresentando um expressivo aumento nos últimos anos, tanto em quantidade quanto na qualidade dos sistemas. Em 2005, Silva (2005), fundamentado em sua pesquisa sobre os graus de participação popular em democracia digital, analisou os portais em operação na rede de 24 capitais brasileiras e constatou que grande parte oferecia apenas serviços informativos. Desta análise, verificou que 87% continham a presença das legislações (leis, estatutos, decretos, portarias etc), possibilidade de inserção de dados pelo usuário de forma a obter uma informação (consulta customizada) e presença de informações genéricas sobre as cidades (econômicas, culturais, turísticas, históricas, entre outras). Já 91% dos portais apresentavam notícias sobre a administração municipal e informações institucionais genéricas (endereço físico e eletrônico, telefones da administração, função de órgãos da administração pública). Por outro lado, nenhum portal apresentou a possiblidade de operação completa de serviço público via rede e apenas 8% apresentavam algum tipo de atendimento on-line instantâneo como, por exemplo, um chat ou a possibilidade de obtenção de serviço público em domicílio com pedido inicial através do site. Por fim, concluiu que grande parte das aplicações encontravam-se nos graus 1 e 2 e que nenhum portal apresentava alguma possibilidade de participação cidadã em nível dos graus 4 e 5.

Conforme a pesquisa, os sistemas existentes no ano de 2005 não propiciavam a participação ativa da população, seja de forma consultiva ou deliberativa. Porém, o avanço do alcance da internet, a melhoria do acesso da população aos recursos de informática, bem como o aprimoramento dos aparatos tecnológicos levaram a um amadurecimento e surgimento de novas iniciativas. Segundo Wildauer, Inaba e Silva (2013), enquanto em 2005 apenas 12% dos domicílios brasileiros tinha acesso à internet, em 2012 esse índice passou para 40% e projeta-se um índice de 50% até o final de 2014. Apesar de ainda estar próximo de atingir a metade dos domicílios brasileiros, garantir o acesso de todo cidadão à internet é uma condição fundamental para o desenvolvimento da democracia digital, possibilitando a comunicação, a rapidez, a busca e disseminação da informação em massa, sendo, portanto, um importante canal de livre acesso e exercício da cidadania.

A popularização da internet está permitindo o surgimento das chamadas “cidades inteligentes”, uma forma de designar as cidades que fazem uso das inovações tecnológicas para facilitar o acesso da população aos serviços públicos e solucionar os problemas da sociedade local. Iniciativas nesta área já podem ser encontradas em algumas cidades brasileiras, as quais procuram desenvolver portais de serviços online que permitem o acesso a diversos serviços públicos, busca de informações, pagamentos de tributos municipais, além de oferecer espaço de ouvidoria online

para que as comunidades interajam com suas prefeituras fazendo solicitações e reinvindicações. Uma das características das "cidades inteligentes" brasileiras é a preocupação com a inclusão digital, uma vez que serviços online são só democráticos de fato quando 100% da população têm acesso a eles (PACHECO et al., 2013).

No entanto, dois projetos se destacam tanto pelo apelo à participação popular quanto pela inovação tecnológica no que se refere a um pioneiro tratamento da informação em prol de uma inteligência coletiva que corresponda aos anseios da comunidade. O primeiro destaque é o portal e-democracia da câmara dos Deputados e o segundo, o portal Gabinete do Governo mantido pelo Governo do Estado do Rio Grande do Sul. Ambas iniciativas serão detalhadas a seguir, principalmente pelo fato do escopo social e tecnológico se aproximar dos objetivos do projeto deste trabalho de mestrado.

O portal e-democracia da Câmara de Deputados foi criado em 2009 e seu objetivo, segundo descrição no próprio site (http://www2.camara.leg.br/), é *"incentivar a participação da sociedade no debate de temas importantes para o país através da internet"*. É classificado como um sistema sócio-tecnológico pois são soluções tecnológicas inovadoras que possibilitam a colaboração popular rápida, eficaz e abrangente. Por este motivo, sua implementação envolve ampla participação interdisciplinar que leva em consideração a interação dos seus componentes tecnológicos e sociais e culmina em um instrumento antropocentrado e não apenas tecnocentrado (MEZARROBA et al., 2013).

O portal é formado por dois espaços de interação denominados Comunidades Legislativas e Espaço livre. No primeiro, o usuário participa de debates a respeito de temas específicos relacionados a projetos de lei que tramitam na câmara. No segundo, o usuário pode criar e participar de fóruns sugerindo temas a serem discutidos. Além disso, há um espaço para compartilhar informações com redes sociais como o Facebook e o twitter, bate-papo, enquetes e uma Wiki da comunidade virtual.

No entanto, a principal colaboração cidadã está no instrumento de votação de ideias e temas que permite conhecer a opinião, anseios e propostas da comunidade a respeito de assuntos que estão em pauta na câmara dos deputados. Esta votação utiliza-se de uma metodologia de *crowdsourcing* cuja proposta é produzir um sistema de seleção de ideias em um universo de grande participação colaborativa. Essa metodologia será detalhada mais adiante.

Como exemplo da participação cidadã através do e-democracia destaca-se o projeto recentemente aproavado pela câmara dos Deputados que criou o Marco Civil da internet. Foi o primeiro projeto de lei de importante impacto social que teve incorporado sugestões dos quase 12 mil internautas que acessaram a comunidade virtual e contribuiram com ideias. Das 374 manifestações selecionadas pelo sistema e colocadas em votação, 6 foram incorporadas pelo relator e passaram a fazer parte do projeto de lei aprovado.

O outro destaque é o portal "Gabinete Digital" desenvolvido pelo Governo do Rio Grande do Sul com o propósito de ser um canal de participação e diálogo entre governo e sociedade. Segundo a descrição encontrada no site (http://gabinetedigital.rs.gov.br/), o objetivo é *"incorporar novas ferramentas de participação, oferecendo diferentes oportunidades ao cidadão de influenciar a gestão pública e exercer maior controle social sobre o Estado"*. Inspirado em projetos de democracia digital tanto no Brasil quanto no exterior, o portal, criado em 2011, já se tornou

foco de pesquisas acadêmicas e recebeu diversos prêmios nacionais e internacionais pelo incentivo de pesquisas e projetos em participação popular, cultura digital, propriedade intelectual e democracia.

Segundo Wu (2013), o sistema está apresentando um novo paradigma para participação cidadã, indo além das abordagens únicas sobre as questões sociais para uma abordagem diversa onde uma grande quantidade de pessoas são ouvidas e atendidas. Wu (2013) descreve o projeto como uma aposta na renovação da democracia não apenas por criar um instrumento de participação cidadã, mas também por ser um sistema, que, por si só, é de apropriação pública já que a plataforma foi desenvolvida com tecnologia aberta e licenças livres.

O Gabinete Digital é um ambiente digital que processa as ideias de um debate público e constrói consenso através da metodologia de *crowdsourcing*, a mesma utilizada no portal e-democracia da Câmara dos deputados citado anteriormente. Elaborado com o objetivo de alcançar a maior participação popular possível, foi desenvolvido em código aberto, o que permitiu a interação com outros especialistas técnicos de forma colaborativa e, inclusive, a replicação do projeto técnico em outros municípios do Brasil. Além disso, o material gráfico, como os vídeos, foram distribuídos em formatos que possibilitam a visualização sem a necessidade de softwares específicos, como por exemplo, em HTML5. A interface é adaptável a dispositivos móveis como tabletes e smartphones.

A proposta do projeto é estabelecer uma relação entre a tecnologia de informação e os usuários de forma que toda informação colaborativa seja base para debates e deliberações políticas e não meramente consultivas, criando um vínculo efetivo entre as esferas pública e política. Sua principal ferramenta colaborativa é denominada "Governador Pergunta" onde os usuários devem responder a uma questão que diz respeito a temas de interesse público. As propostas são recebidas e disponibilizadas para votação e, ao final, os autores das questões selecionadas são convidados a participar de um encontro presencial com o Governador do Estado debatendo o encaminhamento das propostas levantadas. Outras ferramentas também estão disponíveis como o "Governador Responde" onde qualquer cidadão pode enviar um questionamento para o governador, as quais ficam a disposição de todos e são colocadas em votação. A pergunta mais votada do mês é respondida pelo governador em vídeo. Já no "Governo Escuta", são realizadas audiências públicas transmitidas pela internet com a participação do público e especialistas onde os usuários podem enviar questões e sugestões ao vivo.

Um exemplo da participação alcançada e dos resultados obtidos pelo programa está em uma consulta feita a respeito da Saúde no Estado do Rio Grande do Sul. Ao todo foram recebidas 3,3 mil propostas e 360 mil votos das quais foram selecionadas 50 prioridades para a saúde no Estado. Desta, pelo menos seis propostas saíram do papel e se tornaram lei, mostrando que a parceria entre tecnologia e democracia é um caminho eficiente para promover a participação cidadã direta na solução dos problemas sociais.

Tanto o projeto e-democracia da Câmara dos Deputados quanto o projeto Gabinete Digital do Governo do Estado do Rio Grande do Sul exploram uma metodologia de *crowdsourcing* desenvolvida na Universidade de Princeton através do projeto "All Our Ideas". Segundo Brabham (2009) *Crowdsourcing* é uma abordagem que usa o conhecimento, a energia e a criatividade de uma comunidade online para resolver algum problema levantando coletivamente as soluções.

O "All Our Ideas" tem como objetivo desenvolver uma pioneira forma de coleta de dados sociais combinando características de métodos quantitativos e qualitativos. Nascido em 2010, a proposta inicial era criar uma ferramenta que pudesse coletar e priorizar as ideias dos estudantes da Universidade, porém, conforme foi sendo concretizada, converteu-se em um amplo projeto de coleta de dados sociais cujo intuito é ter a amplitude, velocidade e quantificação de uma pesquisa quantitativa enquanto, ao mesmo tempo, permita o acréscimo de novas informações como acontece em uma entrevista. Além disso, propõe ser um projeto colaborativo, desenvolvido em software livre e a disposição da comunidade para uso, revisão e reformulação (fonte: http://allourideas.org).

A plataforma é composta por uma *wiki-survey*, uma página on-line onde é possível criar uma consulta e a torná-la pública. Nesta consulta, cada usuário poderá ou colocar uma ideia sobre o tema do *survey*, reforçando o conjunto de idéias elegíveis, ou votar nas ideias que já estão nesse conjunto, procedimento o qual permite a coleta e a priorização de informações em um único processo. No processo de votação, as ideias são apresentadas ao usuário de forma binária, ou seja, duas a duas ("Qual dessas alternativas você prefere?") escolhidas de forma aleatória dentre todas as disponíveis no conjunto de ideias elegíveis. A cada nova votação, o sistema infere estatisticamente a mais provável ordem de preferência global do conjunto, atribuindo a cada ideia uma nota de 0 a 100 onde quanto maior a nota maior a chance da ideia "vencer" caso seja confrontada durante a votação. Desta forma, é obtido um *ranking* que se altera em tempo real conforme vão ocorrendo as votações e novas inserções (SALGANIK; LEVY, 2012). Wu (2013) destaca três vantagens deste procedimento de votação:

1. A votação em pares evita desvios da expressão genuína da vontade popular por meio de grupos de interesse.

2. A pontuação é baseada na busca da preferência média da opinião pública, o que envolve o desenvolvimento de algoritmos para cálculos estatísticos, criando um campo sempre passível de aprimoramento.

3. As novas ideias inseridas não são prejudicadas pelo andamento da votação uma vez que a priorização é baseada no número de pontos acumulados nas disputas e não no número de votos.

Segundo consta no site do projeto, de 2010 até março de 2014, a metodologia de coleta de dados sociais já foi usada em mais de 4400 consultas públicas com 214500 ideias e 5,3 milhões de votos mostrando tratar-se de um método útil para seleção de dados sociais, porém, segundo Salganik e Levy (2012), a metodologia possui algumas limitações que devem ser cuidadosamente observadas.

A primeira limitação é que, como as ideias são ranqueadas, a tendência é que aquelas que expressam condições mais generalistas sejam as mais representativas dificultando estimar informações adicionais e detalhadas a respeito da opinião pública. Um exemplo desta limitação está em uma consulta feita na cidade de Nova Iorque onde uma das proposições mais votadas foi "tornar a cidade melhor e mais verde". Porém, apesar de expressar um anseio coletivo, não é possível saber o que as pessoas pensam a respeito do que é uma cidade melhor e mais verde.

Nestes casos, sugerem Salganik e Levy (2012), uma vez conhecida as ideias principais pode-se criar novas consultas que coletem co-ideias a respeito de temas mais genéricos, ajudando o pesquisador a compreender melhor as linhas gerais de uma pesquisa. No caso da cidade de Nova Iorque, uma nova *wiki-survey* poderia explorar a o que seria uma cidade "melhor e mais verde" e concluir que refere-se a melhoras no transporte público ou ao uso mais intenso de bicicleta, facilitando uma tomada de decisão política que vá a favor dos anseios populares. Outra limitação apontada é sobre a validação dos resultados, principalmente no que se refere à necessidade de uma validação mais robusta. Por fim, observa que as *wiki-surveys* são mais adequadas para situações em que exista uma pergunta pré-determinada, não sendo uma metodologia útil para coletar informações de ideias díspares.

A proposta do projeto "All our ideas" possui muitas semelhanças com o presente projeto de mestrado. Tanto as *wiki-surveys* quanto o Extrator de Conhecimento Coletivo, principal produto e objeto de estudo desta dissertação, são ferramentas tecnológicas cujo propósito é buscar de forma eficiente a inteligência coletiva através da participação direta, abrangente e em massa dos usuários, tornando possível conhecer os principais anseios de uma comunidade de forma que se fortaleçam os princípios democráticos. De todas as aplicações estudadas e que hoje são o estado da arte em democracia digital, o projeto da Universidade de Princeton é o que mais se aproxima com os objetivos desta dissertação no que se refere à busca do conhecimento coletivo no propósito de desenvolver a democracia. Porém, apesar deste ponto comum, as metodologias utilizadas são de natureza tecnológica diferentes e, neste aspecto, o extrator de conhecimento coletivo possui algumas características que podem ser vantajosas em relação ao projeto estadunidense.

A primeira diz respeito à sua base tecnológica que o aproxima da área de processamento de linguagem natural, possibilitando superar duas limitações do projeto da Princeton: o problema do detalhamento de ideias e a necessidade de perguntas pré-definidas. Uma vez que as pessoas poderão livremente redigir o que pensam a respeito de assuntos genéricos, o extrator de conhecimento coletivo, além de gerar um ranking com as ideias mais representativas, deverá ser capaz de explorar as ideias adjacentes aos temas principais de forma que não haja perda de detalhes e garanta a diversidade de informação. Além disso, mesmo que as manifestações não digam respeito a um mesmo assunto, nas etapas iniciais do processamento da informação será possível inferir sobre os temas mais relevantes antes de chegar às ideias mais representativas, o que o torna também um sistema classificatório. Por fim, a forma como o usuário interage com o sistema inserindo um conteúdo livre sem ter acesso direto ao conteúdo dos outros usuários pode evitar que o mesmo seja induzido por outras opiniões ou deixe de apresentar peculiaridades do seu ponto de vista ao votar em ideias já existentes. No entanto, apesar de algumas vantagens sobre a metodologia de *crowdsourcing*, o Extrator de Conhecimento Coletivo possui seus próprios desafios e limitações. Discutir tais aspectos, tal como sua metodologia e implementação, será o escopo deste projeto.

Por fim, depois desta caminhada pela história da democracia desde a Grécia antiga até os principais instrumentos de democracia digital da atualidade, ficam mais claros os objetivos desta tese, o qual é desenvolver uma ferramenta sócio-tecnológica que resgate valores democráticos antes impraticáveis como, por exemplo, a participação direta e em massa dos cidadãos

nos negócios públicos. A idéia do Extrator de Conhecimento Coletivo aproxima-se do que o filósofo Pierre Lévy denominou de "Ágora Virtual" ou espaço virtual público de participação política e inteligência coletiva onde é possível conhecer os problemas, desejos e anseios de uma comunidade e tomar decisões de forma democrática. Para atingir este objetivo, a proposta é utilizar conhecimentos da área de inteligência artificial, processamento de linguagem natural e redes complexas como fundamentos para desenvolver a ferramenta. Nos dois próximos capítulos serão abordadas duas áreas que fudamentam o projeto: redes complexas e os sumarizadores extrativos.

Capítulo 2

Redes Complexas

Conceber a ideia do que é uma rede é algo muito simples: trata-se de um sistema formado por um conjunto de itens, denominados nós ou vértices, os quais estão conectados segundo algum critério específico através de arestas. Apesar da simplicidade, a ideia de rede é o fundamento para o estudo de uma enorme quantidade de sistemas reais como, por exemplo, sistemas tecnológicos, sociais, biológicos, linguísticos, dentre outros.

Segundo Diestel (2000) o primeiro estudo envolvendo o uso de redes data do século XVIII e ficou conhecido como o "problema das sete pontes de Konigsberg", resolvido em 1736 pelo matemático Leonhard Euler. O problema consistia na possibilidade de atravessar as sete pontes que cruzavam o rio Prególia, na cidade de Konigsberg, de forma que fosse possível passar por todas sem repetir nenhuma. Euler resolveu usando um raciocínio simples: considerou que as pontes seriam arestas e as faixas de terra que elas levavam seriam nós, reduzindo o problema a um grafo conforme apresentado na Figura 2.1.

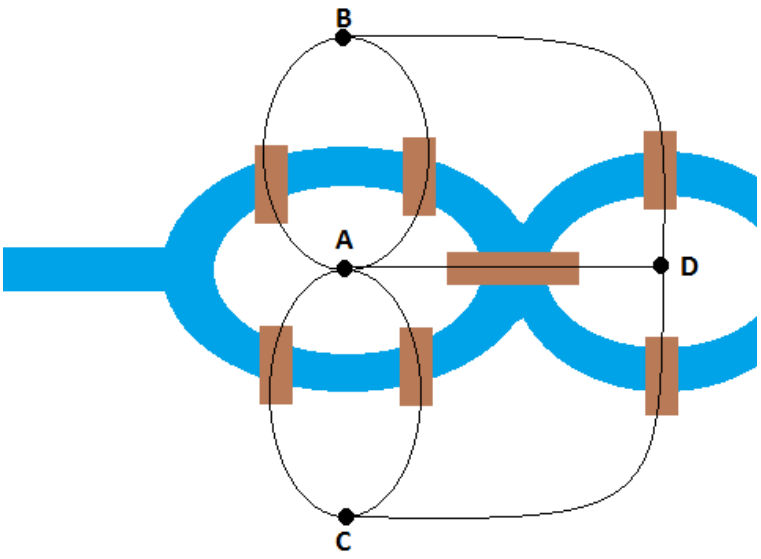

Figura 2.1: Problema das sete pontes de Konigsberg

Desta forma, o problema pôde ser tratado a partir de uma linguagem matemática cujo propósito seria encontrar o caminho que parte de um nó e retorna a ele mesmo passando por todas as arestas apenas uma vez. Euler provou que este caminho não existia ao observar que, para sua existência, cada nó deve ter pelo menos uma entrada e uma saída, com exceção do primeiro e último nó do caminho. Assim, poderia haver, no máximo, apenas dois nós com uma quantidade ímpar de caminhos que chegam ou saem dele. Na linguagem da teoria dos grafos, um caminho euleriano só existe se houver, no máximo, dois nós com grau ímpar, sendo o grau de um vértice o número de arestas ligadas a ele. Como no problema das sete pontes de Konigsberg os quatro nós possuem grau ímpar, pelo teorema enunciado não é possível encontrar um caminho euleriano, logo, o problema não tem solução. A prova de Euler é considerada como o primeiro teorema da teoria dos grafos que, quase 300 anos depois, se tornou a principal linguagem matemática para descrever as propriedades das redes (NEWMAN; BARABASI; WATTS, 2006).

Se o nascimento da teoria dos grafos está relacionado aos estudos de Euler, sua extensão é em grande parte devido aos trabalhos do matemático Paul Erdős. Como no caso de Euler, Erdős se interessou pela matemática dos grafos ao tentar solucionar um problema real: qual é a estrutura das redes sociais? Esse problema, formulado pela primeira vez por Kochen (1989 apud AMARAL; OTTINO, 2004) na década de 50, culminou na definição de um modelo de redes conhecido como redes aleatórias. Dentre algumas importantes propriedades da teoria dos grafos que Erdős definiu está o limiar de percolação (número médio de arestas por nós necessárias para um grafo estar totalmente conectado) e o número médio de arestas intermediárias no caminho mais curto entre quaisquer dois nós de um grafo (AMARAL; OTTINO, 2004).

Outro importante estudo para teoria dos grafos veio do psicólogo Milgram (1967), o qual pesquisou como era a rede de contatos da sociedade estadunidense a partir de um experimento que envolveu o envio de correspondências por parte dos participantes para pessoas desconhecidas. Deste trabalho, concluiu que, em média, havia 6 pessoas entre duas pessoas quaisquer, fenômeno que ficou conhecido como "seis graus de separação" e foi descrito na linguagem de grafos como efeito mundo pequeno, uma das características mais recorrentes em redes do mundo real.

Porém, foi apenas mais recentemente que a teoria dos grafos se aproximou do estudo de sistemas complexos reais. Este novo impulso foi fruto dos trabalhos de Watts e Strogatz (1998) e Barabási, Albert e Jeong (1999). Os primeiros propuseram um novo modelo de redes que descreve a emergência do efeito mundo pequeno em redes simples ao constatar que inúmeros sistemas reais apresentavam este fenômeno. Já Barabási, Albert e Jeong (1999), também analisando algumas redes oriundas de sistemas reais, observaram que alguns poucos nós estavam fortemente conectados enquanto a grande maioria deles possuía um pequeno número de conexões, tal como em uma distribuição que segue a lei da potência (NEWMAN, 2010). Os grafos que apresentaram este comportamento foram denominados redes livres de escala.

Estes estudos não apenas deram origem a novos modelos de rede como também inauguraram um novo campo da ciência denominado Redes Complexas, o qual une diversas áreas científicas como a própria teoria dos grafos (BOLLOBÁS, 1998) e a mecânica estatística (CHANDLER, 1987). No entanto diversas outras áreas se beneficiam destes estudos como a biologia, psicologia, neurociência, linguística, sociologia (ALBERT; BARABÁSI, 2002) e outras cujo objeto de

análise seja um sistema complexo. Os sistemas complexos serão assunto do próximo tópico.

2.1 Sistemas Complexos

Sistemas como a *World Wide Web* (HUBERMAN; ADAMIC, 1999), Internet (PASTOR-SATORRAS; VESPIGNANI, 2007), rede de colaboração entre atores (WATTS; STROGATZ, 1998), rede de citações científicas (REDNER, 1998), rede de contatos sexuais (LILJEROS et al., 2001), sistemas de ruas urbanas (JIANG; CLARAMUNT, 2004), proteínas (MASLOV; SNEPPEN, 2002), redes neurais (BULLMORE; SPORNS, 2009) e cadeias alimentares (MONTOYA; SOLÉ, 2002) compartilham uma característica comum: a topologia de rede. A estes sistemas denominam-se sistemas complexos e têm na teoria dos grafos um importante corpo de conhecimento que pode ser aplicado para sua descrição, análise e entendimento.

Segundo Mitchell (2006), a definição de sistemas complexos não está formalmente descrita na literatura científica porém, informalmente pode ser entendida como um sistema que possui uma topologia de rede composta por componentes relativamente simples, sem nenhum controle central e que exibe um comportamento complexo emergente. Os componentes são ditos relativamente simples por serem entendidos em relação ao comportamento emergente e não a sua individualidade. Por exemplo, uma formiga isoladamente é uma estrutura sofisticada, porém o seu papel é relativamente simples se comparado com o comportamento emergente de uma colônia. Já o comportamento emergente é dito complexo por ser entendido como um comportamento que emerge de um conjunto de ações de agentes simples onde mapear as ações individuais para o comportamento coletivo não é uma tarefa trivial. Desta forma, um sistema complexo é mais do que a soma de suas partes e para caracterizá-lo é necessário conhecer sua topologia de rede, como se realiza o processamento de informação e o quão adaptados estão estes processos no sistema como um todo. Em geral, os sistemas complexos possuem um grande número de componentes que interagem de acordo com regras nem sempre conhecidas e que podem mudar ao longo do tempo. Assim, uma importante característica destes sistemas é a sua capacidade de responder às condições externas e se modificar ao longo do tempo.

Na mesma direção, Amaral e Ottino (2004) também buscam uma definição para sistemas complexos e concluem que:

> *"É um sistema com um grande número de elementos, blocos ou agentes, capazes de interagir uns com os outros e com seu ambiente. Essa interação pode ocorrer tanto com os elementos vizinhos quanto com elementos distantes; os agentes podem ser tanto idênticos quanto diferentes; eles podem se mover no espaço ou ocupar posições fixas e podem estar em um de dois ou múltiplos estados. A característica comum de todos os sistemas complexos é que eles mostram organização sem que qualquer princípio externo organizador seja aplicado."*

O foco de estudos está em descobrir e explicar as leis comuns que levam ao comportamento coletivo emergente. Porém, para isto, Amaral e Ottino (2004) citam três condições que devem ser cuidadosamente observadas:

1. Natureza das unidades: sistemas complexos tipicamente apresentam um grande número

de agentes, porém nem sempre todos possuem estruturas idênticas.

2. Natureza das interações: os agentes interagem fortemente e, frequentemente, de forma não linear. Faz-se necessário conhecer estas interações, a estrutura da rede e os ruídos que a interferem.

3. Natureza da energia inicial: sistemas complexos são tipicamente fora do equilíbrio e suscetíveis a perturbações externas.

Segundo Mitchell (2006), muitos cientistas acreditam que descobrir os princípios gerais que regem um sistema é uma condição essencial para se criar vida e inteligência artificial. Alguns sistemas complexos estão entre as mais fascinantes questões científicas da atualidade como, por exemplo, a emergência da consciência a partir da interação dos neurônios; como seres humanos criam as regras sociais; ou como o DNA organiza os processos no interior das células.

Em relação ao ferramental de estudos dos sistemas complexos, três áreas da ciência se destacam: a dinâmica não linear, a física estatística e a teoria de redes complexas. A primeira área fundamenta-se na teoria do caos para compreender como sistemas simples altamente sensíveis às condições iniciais e cujos comportamentos são difíceis de prever, pois não mantêm memória, produzem saídas complexas. Já a física estatística busca as leis universais que regem o comportamento emergente independente dos detalhes microscópicos de seus agentes. Por fim, a teoria de redes complexas, a qual busca modelar os sistemas reais como grafos e entender o comportamento emergente a partir de métricas e medidas matemáticas sobre a topologia e dinâmica da rede. Esta área tem ganhado importância nos últimos anos devido ao aumento da capacidade de memória e processamento dos computadores, possibilitando estudar de forma eficiente redes formadas por inúmeros agentes (AMARAL; OTTINO, 2004). A teoria de redes complexas é o foco do próximo tópico.

2.2 Teoria das Redes Complexas

A linguagem matemática que descreve as redes complexas é oriunda da teoria dos grafos. Formalmente, uma rede complexa é modelada como um grafo cujos nós representam os agentes ou componentes de um sistema complexo real e as arestas uma relação entre eles. Matematicamente (DIESTEL, 2000), um grafo é um conjunto $G = (V, E)$ tal que $E \in [V]^2$, ou seja, os elementos de E são um subconjunto de 2 a 2 elementos de V. Os elementos de $V = \{v_1, v_2, v_3, \cdots, v_n\}$ são chamados de vértices (ou nós) do grafo G e sua quantidade é representada por N. Os elementos de $E = \{e_1, e_2, \cdots, e_m\}$ são denominados arestas (ou linhas) do grafo G e sua quantidade é representada pela letra M. A forma usual de expressar visualmente um grafo é desenhando um ponto para cada vértice e, entre eles, uma linha representando a aresta, tal como o grafo da figura 2.2.

Grande parte das redes complexas é formada por grafos que possuem apenas uma aresta entre um par de vértices. Porém, em alguns casos, pode ocorrer do grafo possuir mais do que uma aresta entre o mesmo par de vértices, sendo, nestes casos, denominada de multi-aresta.

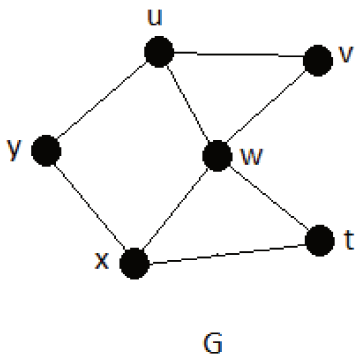

Figura 2.2: Exemplo de um grafo simples.

Também pode ocorrer de uma aresta estar ligada ao mesmo vértice, sendo denominada auto-aresta. Uma rede que não possui nem auto-arestas e nem multi-arestas é chamada de rede simples ou grafo simples, Já uma rede com multi-arestas é chamada de multi-grafo (NEWMAN, 2010). A figura 2.3 ilustra estes casos.

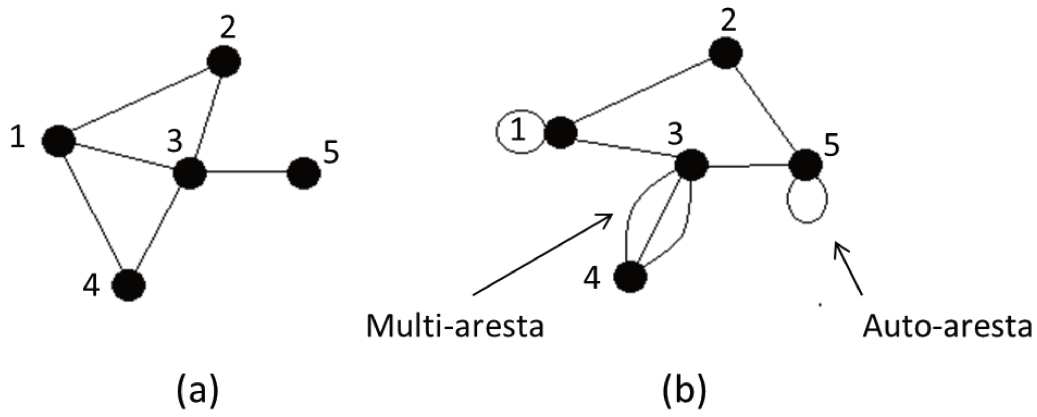

Figura 2.3: (a)Grafo simples; (b)Multi-grafo contendo auto-arestas e multi-arestas.

2.2.1 Representação matemática da rede: Matriz de adjacências

Uma rede pode ser representada matematicamente de diferentes maneiras: graficamente, conforme as figuras 2.2 e 2.3; como uma lista de adjacências; ou como uma matriz de adjacências. Para este estudo, um grafo será representado ora como lista de adjacências e ora como uma matriz de adjacências. Seguem as definições conforme Newman (2010).

Considere um grafo que contém N vértices, chamados de 1, 2, até n e M arestas, chamadas de $1, 2$ até m. Uma aresta entre dois vértices i e j é denotada por (i, j). Desta forma, uma rede pode ser especificada como uma lista de pares (i, j), denominada lista de adjacências. Para a rede da figura 2.3 (a), tem-se a seguinte lista de adjacências: $(1, 2), (1, 4)(2, 3)(3, 1)(3, 4)(3, 5)$.

Apesar de ser uma representação útil para desenhar grafos em computadores, fazer cálculos matemáticos com a lista de adjacências é custoso. Para facilitar este procedimento, a melhor representação de um grafo é a partir de uma matriz de adjacências. Uma matriz de adjacências A de um grafo simples é a matriz de elementos A_{ij} tal que

$$a_{ij} = \begin{cases} 1, & \text{se há uma aresta entre os vértices } i \text{ e } j \\ 0, & \text{caso contrário.} \end{cases} \tag{2.1}$$

No exemplo da figura 2.3 (a), a matriz de adjacências é a seguinte:

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Também é possível representar multi-arestas e auto-arestas usando esta representação. Uma multi-aresta é representada atualizando o valor do elemento A_{ij} pela multiplicidade da aresta. Por exemplo, uma aresta dupla entre os vértices i e j é representada por $A_{ij} = A_{ji} = 2$. Já a auto-aresta é representada pelos elementos da diagonal principal. Por exemplo, uma auto-aresta simples que está no vértice 2 é representada como $A_{22} = 1$. A seguir, serão tratados dois casos especiais de redes: as redes direcionadas e as redes com pesos e depois serão apresentadas as principais propriedades das redes complexas.

2.2.2 Redes com pesos

Muitas vezes as arestas de um grafo são unicamente representadas em uma lista ou matriz de adjacências pela presença ou ausência entre dois vértices. Porém, em algumas situações faz-se necessário representar as arestas como tendo uma força, peso ou valor, geralmente expresso como um número real. Por exemplo, em uma rede que representa a Internet, arestas com pesos poderiam representar a quantidade de dados que flui ao longo de um canal.

Segundo Newman (2010), grafos que possuem arestas com pesos são denominados redes ou grafos com pesos e podem ser representados matematicamente através da matriz de adjacências, atribuindo aos elementos a_{ij} o peso correspondente às ligações como no exemplo a seguir, onde a força da aresta que une os vértices 1 e 2 é o dobro da que une o 1 e 3, o qual, por sua vez, é o dobro da força na aresta entre os nós 2 e 3:

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 0.5 \\ 1 & 0.5 & 0 \end{pmatrix}$$

Os valores das arestas em uma rede com pesos são geralmente números positivos, todavia não há nenhum impedimento teórico para que sejam números negativos. Em redes sociais é comum representar pesos positivos como representação para uma relação cordial entre duas pessoas e pesos negativos para relações não cordiais. Além disso, vale notar que uma rede multi-grafo também pode ser vista como uma rede com pesos com valores inteiros positivos. Tal circunstância pode facilitar a análise das redes em situações específicas.

Os pesos de uma rede podem também representar distâncias. Em uma rede de linhas aéreas, por exemplo, os valores das arestas podem indicar a quantidade de quilômetros entre dois pontos. Entretanto, deve-se tomar o cuidado durante a análise em notar que arestas que representam distância são analiticamente inversas às arestas que representam pesos ou forças. Quando pesos representam distâncias, valores maiores na aresta indicam distanciamento na topologia da rede enquanto que arestas que representam força indicam uma maior proximidade dos nós. Tal observação faz-se necessária principalmente quando se deseja encontrar aglomerações e calcular métricas de centralidade.

2.2.3 Redes direcionadas

Segundo Newman (2010), uma rede direcionada ou grafo direcionado, também chamado de dígrafo, é uma rede na qual cada aresta possui uma direção, ou seja, parte de um vértice em direção a outro. Tais arestas são chamadas de arestas dirigidas e graficamente são representadas como flechas conforme a figura 2.4.

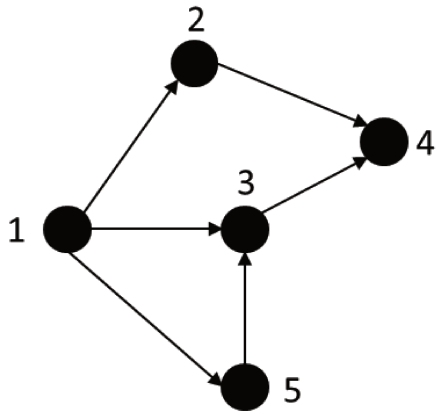

Figura 2.4: Exemplo de uma rede direcionada

A representação de um dígrafo a partir de uma matriz de adjacências é definida como

$$a_{ij} = \begin{cases} 1, & \text{se há uma aresta do vértices } i \text{ para } j \\ 0, & \text{caso contrário.} \end{cases} \tag{2.2}$$

Como exemplo, a matriz de adjacências da rede da figura 2.4 é:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Tal como as redes não direcionadas, os dígrafos também podem conter multi-arestas (ou arestas com pesos) e auto-arestas, as quais são representadas na matriz de adjacências, respectivamente, por elementos com valores maiores que 1 (ou números reais caso sejam pesos) e com elementos não-nulos na diagonal principal.

As redes direcionadas são muito úteis para representar uma ampla gama de sistemas reais tal como a *World Wide Web*, na qual *hiperlinks* indicam a direção de uma *web page* para outra, a cadeia alimentar na qual a energia dirige-se da presa para o predador ou ainda em redes de citações onde citações apontam de um artigo para outro.

2.2.4 Propriedades das Redes Complexas

Uma vez definida a topologia de uma rede, pode-se calcular uma ampla variedade de medidas para quantificar sua estrutura e capturar características particulares de cada topologia. Apesar de muitas dessas métricas terem sido criadas para análise de redes específicas, ela são usadas em diversas aplicações e são a base para caracterizar os modelos de redes complexas a serem vistos mais adiante.

O conhecimento destas medidas permite definir um dos mais importantes conceitos em redes complexas: a centralidade, a qual determina a importância relativa de um vértice no grafo. Existem diferentes métricas para se calcular a centralidade, sendo o grau a mais simples delas. A seguir, serão apresentadas as principais métricas utilizadas para caracterizar as redes complexas e calcular as centralidades segundo definições de Newman (2010) e Ben-Naim, Frauenfelder e Toroczkai (2004).

Grau de um nó

O Grau de um nó é a mais elementar das métricas de uma rede e representa o número de vizinhos que um vértice possui, ou seja, o número de arestas conectadas a ele. Por definição, dada uma matriz de adjacências A de uma rede que contém N nós, o grau k_i de um nó i é dado por

$$k_i = \sum_{j=1}^{N} a_{ij}. \tag{2.3}$$

Sendo o grau uma medida local, muitas vezes é interessante conhecer o grau médio $\langle k \rangle$ de uma rede para caracterizá-la globalmente. Assim

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i. \tag{2.4}$$

Em redes direcionadas, os vértices podem possuir tanto arestas que chegam quanto arestas que saem, sendo possível calcular o grau de entrada (*in-degree*) e o grau de saída (*out-degree*) de cada nó.

Coeficiente de agrupamento

Em muitas redes reais os nós exibem uma tendência a se aglomerar formando grupos na estrutura da rede. A métrica que quantifica essa propriedade é chamada de coeficiente de agrupamento C e reflete a extensão na qual os vizinhos de um nó em particular estão conectados com todos os outros.

Formalmente, o coeficiente de agrupamento é definido por

$$C_i = \frac{2n_i}{k_i(k_i - 1)} \tag{2.5}$$

onde n_i é o número de arestas conectando os k_i vizinhos do nó i para cada outro nó. Caso nenhum vizinho esteja conectado, $C_i = 0$, porém se todos os seus vizinhos estiverem interconectados, $C_i = 1$. Também dada que esta é uma medida local para cada nó, muitas vezes é interessante conhecer o coeficiente de agrupamento médio $\langle C \rangle$ para se caracterizar a rede

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{2.6}$$

Comprimento médio do caminho

Em diversas redes é possível notar a existência de diversos caminhos entre dois nós i e j. Uma importante medida das redes complexas é o tamanho do caminho mais curto entre dois vértices, denominado l_{ij}. O comprimento médio do caminho $\langle l \rangle$ de uma rede é dado pela média de todos os menores caminhos entre dois pares de nós e é definido como

$$\langle l \rangle = \frac{2}{N(N-1)} \sum_{i<j} l_{ij}. \tag{2.7}$$

O menor caminho também desempenha um papel importante na navegabilidade sobre a rede, permitindo acessar de forma rápida determinados nós. Por exemplo, em uma rede de computadores ligados na Internet, o menor caminho torna possível a transferência rápida e econômica de dados entre dois dispositivos. Uma famosa ocorrência desta propriedade está no estudo do psicólogo Stanley Milgran (MILGRAM, 1967) que concluiu que entre dois quaisquer cidadãos estadunidenses havia um comprimento médio do caminho de 6 pessoas. Este fenômeno foi denominado de efeito mundo pequeno e é a principal propriedade de redes reais denominadas redes mundo pequeno. Os modelos de rede serão tema da próxima sessão.

2.2.5 Modelos de Redes Complexas

A aproximação da teoria de grafos ao estudo dos sistemas complexos não apenas forneceu um ferramental matemático para o modelamento e análise de redes reais como também ampliou a própria teoria dos grafos ao propor novos modelos de redes. O primeiro modelo conhecido foi proposto por Erdős e Rényi (1959) e foi denominado de redes aleatórias. Quase 40 anos depois, Watts e Strogatz (1998) ao estudar as redes aleatórias, observaram uma importante característica a qual ficou conhecida como efeito mundo pequeno e formalizaram o modelo de rede mundo pequeno. Por fim, Barabási, Albert e Jeong (1999), aprofundando o estudo das redes mundo pequeno, propuseram o modelo livre de escala. Apesar de existirem outros modelos de redes complexas, estes três representam uma ampla gama de sistemas complexos e, junto com o modelo de redes regulares, serão apresentados a seguir.

Redes regulares

Na teoria dos grafos, redes regulares são aquelas nas quais todos os nós possuem o mesmo grau, conforme é possível observar na figura 2.5:

No caso extremo, onde todos os nós estão conectados entre si, obtém-se uma rede regular completa cujo comprimento médio do caminho é mínimo (igual a 1) e o coeficiente de agrupamento é máximo. Apesar de nem todas as redes reais regulares conhecidas serem completamente acopladas, os valores de l e C tendem a ser elevados, raramente apresentando fenômenos como o efeito mundo pequeno e o comportamento livre de escala.

Embora seja um importante modelo para teoria dos grafos, redes regulares não são comuns na modelagem de sistemas reais. Uma rede completamente acoplada formada por N nós apresenta $N(N-1)/2$ arestas enquanto que a maioria das grandes redes reais apresentam-se esparsas, isto é, não completamente conectadas e seu número de arestas é geralmente da ordem N, ao invés de N^2 (WANG; CHEN, 2003).

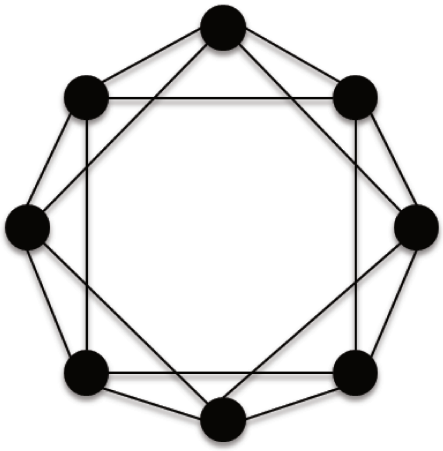

Figura 2.5: Exemplo de uma rede regular.

Modelo de Erdős-Rényi

Antes de 1960, a teoria dos grafos preocupava-se apenas com as propriedades individuais das redes. Porém, neste período, Erdős e Rényi (1959), fundamentado em teorias da probabilidade e na matemática discreta, começaram um sistemático estudo focado nas propriedades estatísticas de conjuntos de redes e não apenas em suas propriedades individuais, o qual culminou no desenvolvimento do modelo conhecido como redes aleatórias.

Segundo West et al. (2001), nesta abordagem, os métodos probabilísticos são usados para mostrar a existência de ligações entre os objetos de uma rede sem que necessariamente ela ocorra, tratando-as não como um objeto de existência real, mais sim como um evento probabilístico. Assim, grafos aleatórios são redes cujos dois nós estão conectados por uma probabilidade p.

Matematicamente, um grafo aleatório $G(N, M)$ é um modelo de rede onde um conjunto específico de parâmetros possui um valor fixo, porém é aleatório em outros parâmetros (NEWMAN, 2010). Um exemplo é uma rede na qual o número de vértices N e arestas M são fixos porém, pares de vértices são aleatoriamente escolhidos para receber as arestas dadas todas as possibilidades de conexão.

Estritamente, o modelo de grafos aleatórios não é definido em termos de uma única rede, mas sim a partir da distribuição de probabilidades sobre todas as possíveis redes, um conceito emprestado da física estatística. Desta forma, uma rede aleatória $G(N, M)$ é corretamente definida como uma distribuição de probabilidades $P(G)$ sobre todos os grafos G nos quais $P(G) = \frac{1}{\Omega}$ para grafos com N vértices e M arestas, onde Ω é o número total de grafos.

Algumas propriedades de grafos aleatórios são facilmente calculáveis como é o caso do grau médio que é dado por $\langle k \rangle = \frac{2M}{N}$. Porém, Erdős e Rényi (1959) observaram que outras propriedades não são tão fáceis de obter e, para solucionar essa questão, descreveram um modelo especial de redes aleatórias denominado modelo Erdős-Rényi ,$G(N, p)$, onde o número de arestas não é mais fixo, mas sim a sua probabilidade de existir entre dois vértices. A definição formal também é dada a partir de um conjunto de grafos, ou seja, de uma distribuição de probabilidades sobre todas as possibilidades de redes e é dada por

$$P(G) = p^M (1-p)^{\binom{N}{2}-M}. \tag{2.8}$$

A partir das definições apresentadas, algumas propriedades dos grafos aleatórios podem ser deduzidas. A seguir serão demonstradas algumas delas.

- Grau médio

Dado um grafo $G(N,p)$ conforme definido anteriormente, seu grau médio $\langle k \rangle$, também denominado c, é dado por

$$\langle k \rangle = c = (N-1)p. \tag{2.9}$$

O cálculo de $\langle k \rangle$ é importante para analisar a estrutura das conexões da rede e seus *hubs*. De modo geral, observa-se desta propriedade:

1. Para $\langle k \rangle < 1$ as redes são formadas por pequenas sub-redes pouco conectadas.
2. Para $\langle k \rangle > 1$, há o aparecimento de uma sub-rede principal onde a maioria dos vértices da rede estão conectadas a seus componentes.
3. Para $\langle k \rangle \geq lnN$, a rede não apresenta ou apresenta um número pequeno de nós isolados.

- Distribuição de grau

Sabendo que um dado vértice em um grafo aleatório é conectado com probabilidade independente para cada um dos outros $N-1$ vértices da rede, a distribuição de grau em uma rede $G(N,p)$ é dada pela distribuição polinomial

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}, \tag{2.10}$$

a qual caracteriza-se como uma distribuição binomial. Porém, para redes muito grandes, ou seja, com elevado valor de N, o grau médio pode ser aproximado para $\langle k \rangle = Np$ e, após algumas simplificações e manipulações algébricas [1], a distribuição binomial torna-se uma distribuição de Poisson, característica exclusiva de redes aleatórias, conforme apresentado a seguir

$$p_k = \frac{(N-1)^k}{k!} p^k e^{-c} = e^{-c\frac{c^k}{k!}}. \tag{2.11}$$

[1] Para maiores detalhes vide Newman (2010), página 402.

- Comprimento médio do caminho

Segundo Newman (2010), o comprimento médio do caminho de uma rede $G(N, p)$ é dado por

$$\langle l \rangle \approx \frac{lnN}{ln\langle k \rangle}. \tag{2.12}$$

O crescimento logarítmico de $\langle l \rangle$ indica que as redes aleatórias estão sujeitas ao efeito mundo pequeno, porém não necessariamente à lei da potência, típico do modelo livre de escala. Estes fenômenos serão apresentados nos próximos tópicos.

O modelo de redes aleatórias, apesar de simples, é importante para o estudo de redes complexas que modelam sistemas cujo conhecimento sobre os mecanismos de conexão entre seus agentes é ausente e só é conhecida a probabilidade de suas ocorrências.

Modelo Mundo Pequeno

Ao estudar a rede de transmissão elétrica do oeste dos Estados Unidos e a distribuição dos neurônios do verme *C. elegans*, Watts e Strogatz (1998) observaram que os componentes dessas redes possuíam uma distância média pequena com valor em torno de 3. Fazendo uma analogia com o fenômeno mundo pequeno observado pelo psicólogo Milgram (1967), definiram o conceito de redes mundo pequeno.

O modelo mundo pequeno nasceu como um extrapolação ao modelo de redes aleatórias, assumindo que redes reais podem apresentar uma topologia que não seja totalmente randômica, mas sim intermediária entre uma rede regular e uma rede aleatória oriunda da movimentação das arestas de um anel regular.

Considerando uma rede regular contendo N vértices e k arestas por vértices ligadas aos seus vizinhos mais próximos, as arestas podem se rearranjar de forma que elas se reconectem aleatoriamente entre os vértices com probabilidade p, indo de uma rede regular ($p \approx 0$) até uma rede aleatória ($p \approx 1$), possibilitando a formação de uma topologia intermediária $0 < p < 1$. A Figura 2.6 ilustra este processo.

Duas métricas são necessárias para caracterizar as redes mundo-pequeno: o comprimento médio do caminho $\langle l \rangle$ e o coeficiente de agrupamento $\langle C \rangle$. Analisando os casos extremos, tem-se que para uma rede regular onde $p = 0$, $\langle l \rangle$ e $\langle C \rangle$ apresentam valores elevados, respectivamente, $\langle l \rangle = \frac{N}{2k} >> 1$ e $\langle C \rangle = \frac{3}{4}$, não havendo, desta forma, o efeito mundo pequeno. Para $p = 1$, o modelo converge para uma rede aleatória com baixo $\langle l \rangle$ e $\langle C \rangle$ com valores aproximados a $\langle l \rangle \approx \frac{lnN}{lnk}$ e $\langle C \rangle \approx \frac{\langle k \rangle}{N}$. Os casos intermediários ilustram as redes mundo pequeno as quais, em geral, apresentam valor elevado de coeficiente de agrupamento e um baixo comprimento do caminho médio com valor aproximado de lnN.

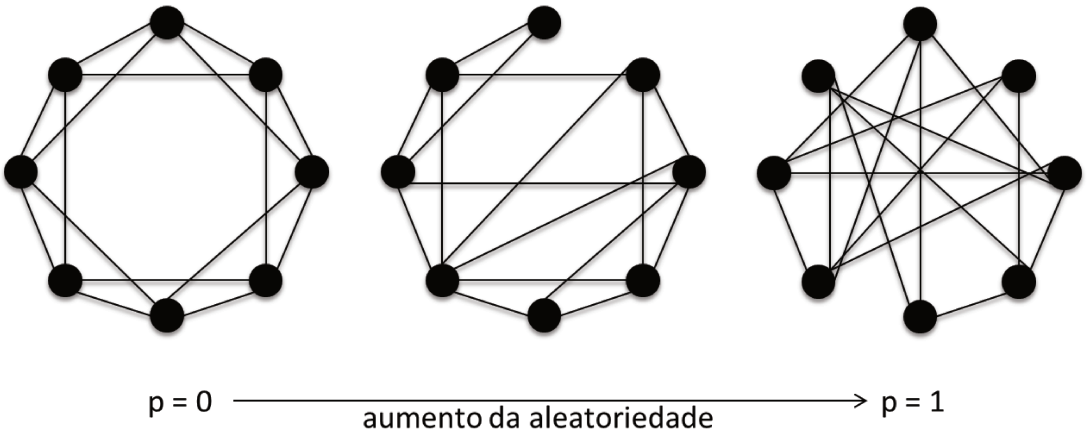

Figura 2.6: Procedimento de formação de redes com N = 8 segundo o modelo mundo pequeno. A primeira rede, onde $p = 0$, é uma rede regular de grau 4. Na segunda rede alguns nós e arestas foram escolhidos para se reconectarem aleatoriamente com probabilidade p. Na terceira rede todos os nós da rede foram reconcetados aleatoriamente, formando uma rede aleatória onde $p = 1$.

Redes Livres de Escala

Ao descrever o modelo de redes mundo pequeno, Watts e Strogatz (1998) não levaram em consideração uma importante propriedade das redes complexas: a distribuição dos graus da rede. Ao estudar alguns sistemas como a *World Wide Web*, Barabási, Albert e Jeong (1999) perceberam que em redes reais alguns poucos nós estão altamente conectados enquanto a maioria possui poucas ligações. Para quantificar este feito, propuseram um novo modelo de redes complexas fundamentado na propriedade da distribuição de graus p_k, o qual ficou conhecido como rede livre de escala.

Em redes de Erdős-Rényi com uma grande quantidade de nós foi observado que a distribuição dos graus se aproxima de uma distribuição de Poisson. Porém, em grande parte das redes reais a distribuição dos graus é altamente enviesada e se decompõe muito mais devagar do que na distribuição poissoniana, apresentando comportamento compatível com a lei da potência com $p_k \approx k^{-\gamma}$. Redes reais como a internet (FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999), redes de chamadas telefônicas (ABELLO; BUCHSBAUM; WESTBROOK, 1998), rede metabólicas (JEONG et al., 2000), a *World Wide Web* (BRODER et al., 2000), foram estudadas e encontrado um expoente γ sempre entre 2 e 3, o qual representa na prática uma distribuição onde poucos nós possuem muitas ligações e muitos nós estão pouco conectados.

As redes livres de escala fundamentam-se em dois pressupostos: possuem uma dinâmica típica de crescimento tal como os outros dois modelos de rede; a dinâmica de crescimento baseia-se em conexões preferenciais, o qual garante uma rede bastante robusta e resistente a falhas aleatórias em um nó, porém vulnerável caso falhas atinjam os nós de elevado grau podendo fazer com que a rede inteira rapidamente se desfaça. Segundo Barabási, Albert e Jeong (1999) uma rede livre de escala emerge em um contexto onde cada novo nó conecta-se preferencialmente aos nós mais conectados já existentes na rede, conforme a equação

$$p_i(n+1) = \frac{k_i(n)}{\sum_{i=n_0+1}^{n} k_i(n)} \tag{2.13}$$

onde n é o número de nós adicionados na rede, n_0 a quantidade inicial de nós no tempo zero, k_i é o grau do nó i e $p_i(n+1)$ é a probabilidade de um novo nó, colocado no tempo $N+1$, ligar ao nó i.

Uma característica importante das redes livre de escala é a existência de *hubs*, ou seja, nós com elevado grau ou elevada centralidade. Esta propriedade torna o modelo especialmente importante para o estudo de redes de informação e conhecimento uma vez que os *hubs* podem ser interpretados como núcleos de onde parte a informação e rapidamente alcança qualquer outro ponto da rede.

Em relação as suas principais propriedades, as redes livres de escala podem ser analisadas a partir do seu comprimento médio do caminho. Bollobás e Riordan (2004) estudaram esta métrica e observaram que para $M = 1$ seu valor se aproximava de $ln\,\langle N\rangle$. Para valores maiores de M, se aproxima a $\frac{ln(N)}{ln(ln(N))}$, valor significativamente menor do que o $\langle l\rangle$ das redes aleatória e mundo pequeno. A eficiência da topologia livre de escala e a existência de um mecanismo simples de emergência desta topologia levaram muitos pesquisadores a acreditarem na onipresença do modelo em redes reais. Apesar de pesquisas mostrarem que nem todas as redes apresentam o comportamento livre de escala, com algumas restrições é possível encontrá-lo (AMARAL; OTTINO, 2004). Além disso, as redes livres de escala também podem ser entendidas como um subconjunto de todas as redes mundo pequeno por dois motivos:

1. A distância média entre os nós na rede aumenta de forma extremamente lenta com o aumento do tamanho da rede; e

2. O coeficiente de aglomeração é maior do que das redes aleatórias.

Redes Complexas e Processamento de Linguagem Natural

Apesar de ser uma teoria que se propõe a analisar o comportamento dos mais diferentes sistemas reais, as redes complexas têm seus estudos focados principalmente em sistemas tecnológicos, sociais e biológicos. Em relação a área de processamento de linguagem natural, alguns estudos têm aparecido a partir do início dos anos 2000 e estão voltados para a modelagem de textos segundo suas propriedades sintáticas e semânticas e para análise topológica de características subjetivas, informações e conhecimento.

Um dos primeiros estudos foi publicado por Cancho e Solé (2001) e o objetivo foi mostrar que a construção de sentenças em linguagem humana não é um processo aleatório, mas sim reflete toda uma organização oriunda do desenvolvimento evolutivo e do uso histórico das estruturas e combinações lexicais. Para comprovar essa hipótese, foram construídas redes de co-ocorrência de palavras onde cada palavra é um nó e as arestas são direcionadas a palavra que ocorre imediatamente em seguida e analisadas suas estruturas. Deste estudo, duas características importantes de redes complexas foram observadas: (1) o efeito mundo pequeno cuja distância média entre os nós da rede situou-se entre 2 e 3; (2) uma distribuição de graus livre de escala,

sendo visíveis os efeitos sobre a estrutura da rede ao desconectar nós altamente conectados. A descoberta de que a construção linguística pode ser modelada como uma rede complexa abre caminhos para diversos estudos no campo da linguagem e do conhecimento.

Diversas outras redes foram propostas para estudar a linguagem. Sigman e Cecchi (2002) analisaram a estrutura da *Wordnet* (MILLER, 1995), uma base de conhecimento da língua inglesa construída de forma colaborativa que contém todos os lexemas e seus possíveis significados, com o objetivo de compreender a organização global dos lexemas na língua inglesa. Para isto, foi modelada uma rede semântica onde os nós eram as palavras e as arestas indicavam a existência de uma relação semântica entre elas. Duas conclusões foram tiradas deste estudo: (1) a *Wordnet* se comporta como um sistema auto-organizado, apresentando o efeito mundo pequeno e o comportamento livre de escala. (2) A polissemia (uma palavra possuir vários significados) é uma característica importante da língua para que a estrutura de rede apresente o efeito mundo pequeno e uma distribuição que segue a lei da potência. Esses achados indicam que a polissemia pode ser crucial para o pensamento metafórico e generalizações. Na mesma direção, Motter et al. (2002) estudaram a relação entre as palavras e seus significados usando como banco de dados um dicionário de sinônimos e também observaram que a rede formada apresentava o efeito mundo pequeno e o comportamento livre de escala, concluindo que tais resultados são importantes para o estudo da estrutura e evolução das línguas como também para as ciências cognitivas.

Indo além das propostas de modelamento de sistemas linguísticos, o cálculo de métricas sobre estas redes pode trazer uma infinidade de informações a respeito de características subjetivas e conhecimento dos conteúdos textuais. Antiqueira et al. (2007a) estudaram a relação entre a qualidade de um texto e as métricas de redes complexas. Para isto, analisaram textos escritos por estudantes do ensino médio cuja qualidade foi avaliada por um banco de juízes e estabeleceram correlações entre estas notas e métricas da rede. Concluíram que coesão e coerência foram os atributos que melhor apresentaram correlação entre as notas e as métricas, em especial, perceberam que conforme o grau e o coeficiente de agrupamento aumentam, a qualidade do texto tende a diminuir. Estudo semelhante foi feito na área de atribuição automática de autoria, onde as métricas de redes complexas foram usadas para detectar características de estilo de um texto. Os resultados obtidos foram satisfatórios com a vantagem de não se precisar de textos longos de um autor para se detectar suas características (ANTIQUEIRA et al., 2007b).

Outro estudo na área e que se aproxima da mineração de dados é o de Antiqueira et al. (2009) o qual propõe o desenvolvimento de um sumarizador extrativo a partir de uma rede complexa onde os nós são parágrafos de um texto e as arestas indicam a quantidade de palavras semelhantes entre cada parágrafo. Este trabalho, tal como um panorama da sumarização extrativa de textos será o tema do próximo capítulo.

Capítulo 3

Sumarizadores Extrativos

A tecnologia ampliou imensamente a capacidade de armazenamento de informação nas mais diversas formas: textos, imagens, sons, atributos especiais, entre outras. O alto volume e a rápida e intensa taxa de crescimento destes bancos de dados ultrapassam em muito a capacidade humana em analisar, interpretar, organizar e buscar a informação neles contida, gerando a necessidade de metodologias e instrumentos eficazes para manipular esse montante de dados (FAYYAD et al., 1996).

É neste panorama que surge a mineração de dados, a qual consiste em processos de descoberta de conhecimento a partir de uma grande quantidade de dados, abarcando desde o preparo da informação para o posterior processamento até a análise e interpretação dos resultados, passando por processos de busca de padrões, segmentação, classificação, agrupamento, generalização etc. Segundo Fayyad et al. (1996), a mineração de dados faz parte de uma abordagem conhecida como *Knowledge Discovery in Database* (KDD), a qual consiste em processos de extração de informação relevante ou padrões nos dados de grandes bancos de dados que sejam não-triviais, implícitos e potencialmente úteis. Tipicamente, é descrito em cinco etapas conforme apresentado na Figura 3.1.

Existem diversas metodologias para tratar a mineração de dados. Weiss (1998) propõe classificá-la em 4 grandes campos: não-paramétrica, lógica, estimadores estatísticos e baseada em aprendizagem. Dentre as metodologias consolidadas na literatura, destaca-se a Análise Exploratória de Dados (DUBES; JAIN, 1980), Análise de Agrupamentos (ANDRITSOS, 2002), Classificação Automática (BEITZEL et al., 2007), Reconhecimento de Padrões (BISHOP, 1995) e Aprendizado de Máquina (SEBASTIANI, 2002). Em especial na área de processamento de linguagem natural, destacam-se o *PageRank* (PAGE et al., 1999), *Latent Semantic Analysis* (LSA) (DUMAIS, 2004) e técnicas de *Trend Detection* (ETD) (MATHIOUDAKIS; KOUDAS, 2010) (KONTOSTATHIS et al., 2004).

Uma importante aplicação dos métodos KDD está na Sumarização Automática de Textos. Segundo Jones (2007), um sumário nada mais é do que a redução de um texto-fonte a partir de técnicas de seleção e\ou generalização das informações mais importantes. Ao processo de produção de um sumário denomina-se sumarização e este envolve três etapas:

1. Interpretação: texto-fonte é representado segundo alguma abordagem específica

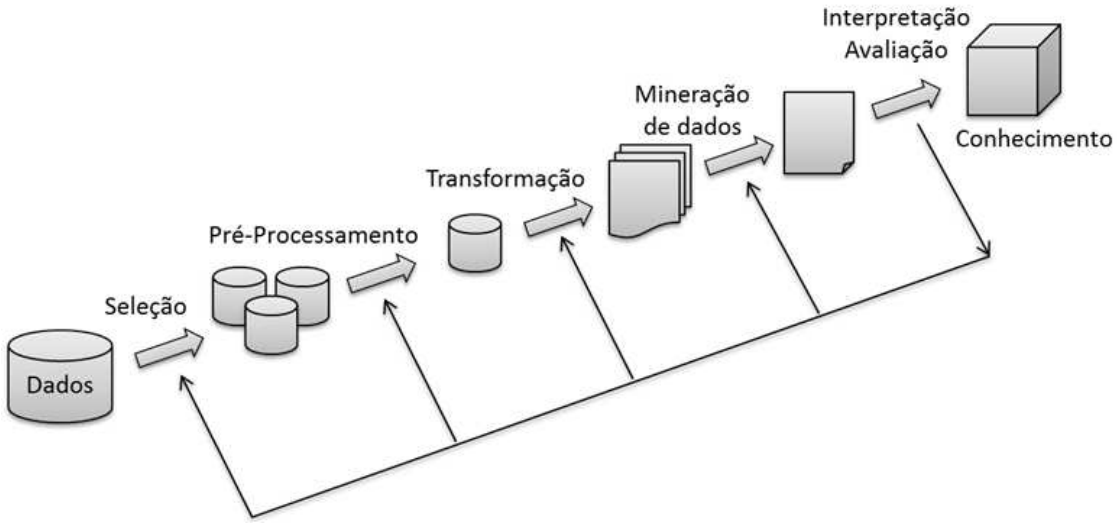

Figura 3.1: Cinco etapas do processo de mineração de dados (FAYYAD et al., 1996)

.

2. transformação: representação do texto-fonte é processada e transformada em uma representação do sumário

3. Geração: representação do sumário é re-interpretado na forma de texto.

As técnicas de sumarização de textos dividem-se em dois grupos segundo a abordagem teórica: técnicas superficiais e técnicas profundas. Na abordagem superficial, métodos estatísticos são utilizados para geração do sumário sem que conhecimento linguístico seja empregado no processamento do texto. No geral, os sumários são produzidos a partir da seleção e justaposição das sentenças do texto original segundo algum critério específico. Tal técnica também é conhecida como sumarização extrativa. Já na abordagem profunda, conhecimento linguístico é amplamente utilizado para construção do sumário, principalmente no que diz respeito a como um texto é organizado. Busca-se produzir uma representação semântica baseado nas relações entre segmentos do texto. Paráfrases, especializações, generalizações e rearranjos são utilizados nas informações selecionadas para compor o sumário final (GUPTA; LEHAL, 2010).

Embora a abordagem superficial esteja mais suscetível a erros como problemas de coesão e ausência de referentes anafóricos, ela é mais simples de ser implementada. A geração de extratos é menos custosa do que a construção de resumos uma vez que, neste último caso, recursos sofisticados de análise são necessários para inferir o significado das sentenças e prover generalizações. Além disso, reproduzir a maneira humana de gerar resumos não é uma atividade simples já que nem sempre é uma produção formalizada e racionalizada e muito ainda se debate sobre processos cognitivos que estão por detrás da geração do conhecimento. Por estes motivos, uma maior atenção tem sido dada para sumarização automática que adota a abordagem superficial (MANI, 2001).

Diversos métodos de mineração de dados podem ser utilizados na sumarização extrativa. Algumas técnicas têm se destacado envolvendo a aprendizagem de máquina: o algoritmo *Support Vector Machines* (SVM) (HIRAO et al., 2002) e *Weighted Probability Distribution Voting*

(HALTEREN, 2002). Outras envolvem o uso de redes neurais na seleção das sentenças mais importantes (KAIKHAH, 2004), lógica fuzzy (SUANMALI; BINWAHLAN; SALIM, 2009), e o método de análise semântica latente (KIREYEV, 2008). Já em relação as abordagens estatísticas, destaca-se o método *Term Frequency-Inverse Document Frequency* (TF-IDF) (RAMOS, 2003).

Porém, a abordagem de maior relevância para o desenvolvimento do Extrator de Conhecimento Coletivo são os métodos baseados em grafos, uma vez que a proposta é modelar o banco de dados como uma rede complexa. Os métodos baseados em grafos visam representar o texto-fonte como uma rede onde sentenças são os nós e as arestas representam algum atributo comum, seja uma relação semântica ou simplesmente a ocorrência de palavras comuns. Uma vez modelado o texto como uma rede, o cálculo de métricas permite selecionar e ranquear os nós, tornando possível a seleção dos parágrafos para compor o sumário. Dentre as técnicas que usam esta abordagem está o LexRank (ERKAN; RADEV, 2004), o algoritmo de *Kleinberg´s HITS* (KLEINBERG, 1999) e o PageRank do Google (PAGE et al., 1999). Em relação ao Extrator de Conhecimento Coletivo, proposta desta dissertação, algumas técnicas desta abordagem inspiraram a criação da arquitetura. A seguir quatro delas serão apresentadas mais detalhadamente.

3.1 Sumarização Automática de textos e Redes Mundo Pequeno

As características mundo pequeno foram observadas por Watts e Strogatz (1998) durante o estudo de redes do mundo real: a Internet, redes sociais, conexões cerebrais, entre outras. Percebeu-se que nestas redes havia um alto grau de aglomeração, um pequeno número de arestas e uma pequena distância média entre os nós. Baseados neste fenômeno, Balinsky, Balinsky e Simske (2011b) apresentam uma proposta de sumarizador extrativo fundamentado na teoria das redes complexas.

A proposta de construir um sumarizador fundamentado nessa abordagem está na premissa de que durante a escrita de um texto, por ser um trabalho humano, as ideias principais e os conceitos são organizados de forma similar à topologia de uma rede biológica. A técnica de sumarização consiste primeiramente em representar os textos como grafos (redes) cujos nós são as sentenças (ou parágrafos). As arestas são usadas para representar as relações entre os pares de nós. O desafio está exatamente em definir parâmetros para tais relações de forma que a rede apresente a característica mundo pequeno.

Para atingir este objetivo, deve-se definir uma "função ranqueadora" (a qual irá pontuar os nós da rede conforme sua importância) que seja capaz de de extrair um pequeno número de sentenças importantes e remover um grande número de sentenças menos importantes, preservando a estrutura do documento.

Neste trabalho, a função ranqueadora é construída a partir do *Princípio de Helmholtz*. Oriundo da teoria gestáltica da percepção humana, é uma função baseada em argumentos da física estatística e comportamentos assintóticos das distribuições normais, cujo objetivo é

descobrir as palavras mais significativas em um texto. Tais palavras estarão num conjunto denominado *MeaningfulSet(E)*. Sua descrição matemática é dada a seguir [1].

Seja D um texto, P uma parte deste texto (por exemplo, um parágrafo) e w uma palavra. Se a palavra w aparece m vezes em P e K vezes em D, então define-se o número de falsos alarmes (NFA) em função de w, P e D pela seguinte expressão

$$NFA = \binom{K}{m} \frac{1}{N^{m-1}}, m > 0 \tag{3.1}$$

onde

$$\binom{K}{m} = \frac{K!}{m!(K-m)!} \tag{3.2}$$

Em 3.1, N é $[\frac{L}{B}]$ onde L é o tamanho do documento D em palavras, B é o tamanho de P em palavras e $[\frac{L}{B}]$ denota apenas a parte inteira de $\frac{L}{B}$. Para medir a importância da palavra w em P é usada a seguinte expressão

$$Meaning(w, P, D) = -\frac{1}{m} log_{10}[NFA(w, P, D)]. \tag{3.3}$$

Finalmente, dado um documento dividido em parágrafos, o $MeaningfulSet(E)$ é definido como o conjunto de palavras

$$Meaning(w, D) > E \tag{3.4}$$

onde $Meaning(w, D)$ é o máximo de $Meaning(w, P, D)$ sobre todos os parágrafos P. Para um valor E, positivo e suficientemente grande, o conjunto $MeaningfulSet(E)$ é vazio. Para um valor de E negativo e próximo de zero, o conjunto $MeaningfulSet(E)$ conterá todas as palavras de D. Para documentos reais, foi observado que o tamanho de $MeaningfulSet(E)$ tem um queda acentuada no número de palavras quando o valor de E está próximo de zero. Uma vez sabido que o valor de E define o tamanho do conjunto de palavras mais significativas do texto ($MeaningfulSet(E)$), o seguinte procedimento é adotado na busca da rede mundo pequeno:

[1] A descrição matemática apresentada se utiliza de alguns conceitos e definições, como por exemplo o significado de *NFA* e *Meaning*, que não serão detalhados nesta dissertação. Porém, podem ser consultados no trabalho de Balinsky, Balinsky e Simske (2011a).

1. Determinado um valor de E, é encontrado o conjunto $MeaningfulSet(E)$.

2. Duas sentenças ou parágrafos do texto são conectados se, e somente se, eles tiverem ao menos uma palavra do $MeaningfulSet(E)$ em comum ou se eles forem um par de sequências consecutivas.

3. Testa-se, então, se a rede formada apresenta o fenômeno mundo pequeno. Caso não apresente, altera-se o valor de E, constrói-se um novo conjunto $MeaningfulSet(E)$ e volta-se para o passo 2 até que seja formada uma rede mundo pequeno.

Se o $MeaningfulSet(E)$ possuir poucas palavras, aparecerão apenas relações locais e o grafo ficará semelhante a uma rede regular. Se, no entanto, o $MeaningfulSet(E)$ possuir muitas palavras, o grafo ficará parecido com uma rede aleatória com muitas arestas. O número de palavras deve ser cuidadosamente escolhido para que o grafo apresente características mundo pequeno. Para testar se a topologia da rede construída apresenta o fenômeno mundo pequeno, duas métricas são calculadas: Comprimento médio do caminho mínimo, já definido no capítulo anterior, e a Transitividade (C), definida a seguir.

Sejam u, v e w nós de uma rede complexa. Se u está conectado com v e v está conectado com w, então existe um caminho uvw de duas arestas no grafo. Se u também está conectado com w, o caminho é um triângulo. A transitividade é então definida como a fração de caminhos de tamanho 2 na rede que forma um triângulo

$$C = \frac{(\text{número de triângulos}) \cdot 3}{(\text{número de triplas conectadas})} \tag{3.5}$$

onde uma "tripla conectada" significa três nós u, v e w com arestas (u, v) e (v, w). São esperados em uma rede mundo pequeno baixos valores de L e altos valores de C.

Uma vez encontrado o valor de E que leve a um pequeno L e um grande C, ou seja, um alcance correto de E de forma a se construir uma rede mundo pequeno, as sentenças que farão parte do sumário são definidas da seguinte forma:

1. Escolha uma medida de centralidade para uma rede mundo pequeno.

2. Verificar se para tal medida de centralidade existe uma grande variedade de valores e uma distribuição *heavy-tail* [2]. Essa distribuição definirá a pontuação das sentenças.

3. Baseado na medida de centralidade, o sumário será composto ou pelas sentenças de maior pontuação ou, caso deseje maior coerência, pelo caminho de sentenças que possuem maior pontuação.

Balinsky, Balinsky e Simske (2011b) escolheram a medida de centralidade Grau para os testes com o sumarizador. Quanto mais conexões possuir o nó, maior será seu grau. Assim,

[2] Também conhecida como cauda longa, é uma característica de distribuições estatísticas como a lei da potência (definida no capítulo 2) onde há uma grande quantidade de dados em baixa frequência arrastando-se em uma longa cauda. Sua definição matemática detalhada pode ser vista em Sigman (1999).

quanto mais conectado estava o nó, maior era sua pontuação. Os sumários gerados por esta técnica foram avaliados por um conjunto de juízes de prova que os avaliaram como "muito bons".

3.2 Algoritmo CN-Sum

Antiqueira et al. (2009) apresentam uma proposta de um sumarizador automático extrativo, denominado *CN-Sum*, que utiliza conceitos e métricas de redes complexas para selecionar sentenças de um texto-fonte. Os nós da rede representam as sentenças do texto e as arestas são colocadas conforme o número de palavras comuns entre as sentenças. Foram testadas 14 métricas de redes complexas para pontuar e selecionar as frases e gerar o sumário. O algoritmo do *CN-Sum* constrói uma rede simples que se baseia na coesão lexical: o texto-fonte é dividido em sentenças onde cada uma representa um nó da rede. Uma aresta é adicionada entre dois nós se as sentenças correspondentes tiverem ao menos uma palavra em comum (isto é, caso ocorra repetição lexical). Cada sentença é previamente preparada retirando-se as *stop-words* (palavras sem significado semântico como artigos e conjunções), os verbos e lematizando (colocando na forma canônica) os substantivos. Antiqueira et al. (2009) observaram que incluir os verbos não melhora o desempenho do sumarizador extrativo.

Em resumo, o método proposto *CN-Sum* consiste em 4 etapas:

1. Pré-processamento: as sentenças são identificadas, os verbos e *stop-words* são retirados e os substantivos são lematizados.

2. Modelagem: o texto resultante é mapeado em forma de matrizes de pesos e adjacências.

3. Teste das métricas: As matrizes são utilizadas para calcular 14 métricas de redes complexas que irão definir uma pontuação para cada nó.

4. Seleção: os primeiros n nós melhor pontuados são escolhidos para formarem o sumário. O número de sentenças que vão compor o sumário é definido pela taxa de compressão escolhida pelo usuário.

No processo de modelagem da rede, são utilizadas duas matrizes para representar o texto: a matriz de adjacências (A) e a matriz de pesos (W). Segue abaixo a descrição matemática da modelagem:

Um grafo não direcional ou rede G de N nós e M arestas pode ser representado por uma matriz de adjacências A, simétrica e de ordem N^2, cujos elementos a_{ij} e a_{ji} são iguais a 1 se há uma aresta entre os nós i e j, ou igual a 0 caso contrário. Tal como explicado anteriormente, uma aresta existe se há uma co-ocorrência de um mesmo substantivo lematizado entre duas sentenças. Se uma aresta tem um "rótulo numérico", isso significa que ela possui um peso. Assim, outra matriz pode ser empregada, chamada de matriz de pesos W, a qual neste caso é definida como a seguir:

Seja $P_i = p_1, p_2, .., p_{n_i}$ o conjunto de n_i substantivos lematizados da *i-ésima* sentença do texto-fonte. O peso w_{ij} (ou w_{ji}) da aresta que une as sentenças i e j é o número de co-ocorrência entre elas, isto é, $w_{ij} = w_{ji} = |P_i \cap P_j|$. Se $w_{ij} = 0$, nenhuma aresta existe entre os

nós i e j. Os pesos w_{ij} são, portanto, elementos da matriz simétrica W de ordem N^2, a qual representa uma rede não direcionada de pesos de um dado texto-fonte.

Diversos experimentos foram feitos testando as 14 métricas de redes complexas para seleção de sentenças, comparando os resultados com outros sumarizadores e utilizando a técnica ROUGE de avaliação de sumários. As métricas que apresentaram os melhores resultados foram respectivamente: as baseadas em Graus, Caminhos Curtos, *D-Rings* e *K-cores*. Na comparação com outros sumarizadores, algumas métricas fizeram com que o *CN-Sum* apresentasse resultados comparáveis ao de sumarizadores que utilizam conhecimento linguístico profundo. Por fim, Antiqueira et al. (2009) concluem que o uso de redes complexas para representar textos é adequado para a sumarização automática, uma vez que a escolha correta das métricas permite capturar as principais ideias de um texto.

3.3 Algoritmos baseados em Grafos

Um dos desafios da sumarização extrativa é determinar a importância das sentenças, classificá-las e selecioná-las para compor o sumário. Thakkar, Dharaskar e Chandak (2010) apresentam dois métodos de extração de sentenças baseados em redes complexas: Algoritmos de ranqueamento e Algoritmos de Caminho mínimo.

3.3.1 Algoritmos de ranqueamento

O Algoritmo de ranqueamento baseado em grafos é uma heurística cujo objetivo é decidir a importância de um vértice em um grafo baseado na estrutura do mesmo. Para que seja possível aplicá-lo, é necessário, primeiramente, representar o documento como uma rede interconectando palavras ou outras entidades do texto (sentenças, parágrafos, etc.) com relações significativas, que podem ser semânticas, lexicais, sintáticas ou de outra natureza. Uma vez definida como será a representação do texto, o algoritmo de ranqueamento baseado em grafos consiste na execução dos seguintes passos:

1. Identificar as entidades textuais que irão compor os vértices.
2. Identificar as relações entre os vértices e adicionar as arestas, as quais podem ser direcionadas ou não direcionadas e podem contar um peso ou não.
3. Aplicar um método que dê uma pontuação aos vértices.
4. Escolher os vértices baseado na sua pontuação final.

Um exemplo de aplicação é o algoritmo *TextRank* (MIHALCEA; TARAU, 2004 apud THAKKAR; DHARASKAR; CHANDAK, 2010), onde os vértices do grafo são sentenças pré-definidas do texto que são conectadas segundo a métrica de redes complexas chamada similaridade. Cada sentença é formada por um conjunto de N palavras. Dada duas sentenças S_i e S_j, onde cada sentença é representada por um conjunto de N_i palavras que compõem a sentença:

$S_i = W_1^i, W_2^1, ..., W_{N_i}^i$, a similaridade entre S_i e S_j é definida como

$$Similaridade(S_i, S_j) = \frac{|W_k|W_k \in S_i \& W_k \in S_j}{Log|S_i| + Log|S_j|}. \tag{3.6}$$

O resultado é um grafo altamente conectado, com um peso associado a cada aresta (valor da similaridade) indicando a força da ligação estabelecida entre vários pares de sentenças no texto. Uma vez construído o grafo, a pontuação dos vértices é dada por meio do Algoritmo de Passo Aleatório (*Random Walk*) que utiliza os valores da similaridade para decidir a pontuação dos vértices. Calculadas as pontuações, são escolhidas as sentenças de maior pontuação para compor o sumário. O número de sentenças escolhidas será de acordo com a taxa de compressão definida pelo usuário.

3.3.2 Algoritmo de caminho mínimo

Uma das dificuldades da sumarização extrativa é construir um sumário que seja agradável de ler, uma vez que eles são formados por diferentes partes retiradas do documento original sem se preocupar com a coesão e lógica da leitura. A ideia do algoritmo de caminho mínimo é tentar melhorar a coesão do sumário formando um caminho de sentenças baseado na similaridade entre elas. Tal como no *TextRank*, inicialmente o texto é dividido em sentenças, as quais serão os vértices do grafo. Uma aresta é colocada caso haja uma similaridade entre as sentenças. A similaridade nada mais é do que possuir ao menos uma palavra igual. Além disso, todas as sentenças possuem uma aresta para a sentença seguinte do texto. O próximo passo é colocar "custos" nas arestas. Quanto mais similares duas sentenças são (ou seja, quanto mais palavras em comum elas possuem), menor o custo de sua aresta. Quanto mais distantes as sentenças estão no texto original, maior é o custo da aresta. Para favorecer a inclusão de sentenças "interessantes", todas aquelas que são tidas como relevantes para o sumário de acordo com métodos de sumarização clássicos têm o custo das arestas que levam a elas abaixado.

O custo de uma aresta que liga a sentença (nó) S_i à sentença S_j é calculado como

$$Custo(i, j) = \frac{(i-j)^2}{\text{Número de palavras comuns}(i, j) \cdot peso_j} \tag{3.7}$$

onde o peso é calculado como

$$peso_j = (1+\text{Número de palavras comuns}(texto, j)) \cdot \frac{1 + \sum w \in S_j \cdot Tf(w)}{\sum w \in text \cdot Tf(w)} \cdot early(j) \cdot \sqrt{1 + |aresta_j|} \tag{3.8}$$

onde $early(j)$ é 2 se $j < 10$ e 1 caso contrário. $Tf(w)$ é a frequência com que a palavra w aparece no texto todo.

Uma vez que a similaridade está baseada no número de palavras em comum entre duas sentenças, sentenças maiores possuem grandes chances de serem similares a outras sentenças. Favorecer longas sentenças é frequentemente bom para um texto mais coeso. Sumários com muitas sentenças pequenas têm uma maior chance de mudanças abruptas e quebra de coesão.

Construído o grafo e calculado o custo das arestas, o sumário é gerado considerando o menor caminho que se inicia com a primeira sentença do texto original e termina na última sentença.

Os N caminhos mais curtos são encontrados partindo do vértice (sentença) inicial e caminhando por todas as arestas, somando seus custos e armazenando numa fila os "valores prioritários", onde o valor prioritário é o custo total do caminho. O caminho com menor custo é então examinado e se ele não termina no nó final, todos os caminhos que possuem este caminho e contêm uma aresta a mais são colocados na fila dos valores prioritários. Caminhos com *looping* são descartados. Sempre que o caminho de menor custo terminar no nó final, um caminho mais curto é encontrado. Esta busca repete-se até que sejam encontrados os N caminhos mais curtos.

As vantagens deste método são a facilidade de implementação e a independência do idioma. Porém, ele pode ser custoso, pois não há nenhum pré-processamento linguístico (remoção de *stop-word*s, entre outros) e não há garantias de convergência do algoritmo.

O uso de métodos gráficos e redes complexas no desenvolvimento de sumarizadores automáticos de texto tem apresentado resultados satisfatórios, principalmente nas técnicas de sumarização extrativa. Alguns destes fundamentos foram base para o desenvolvimento da arquitetura do ECC, principalmente no que diz respeito à escolha da modelagem dos textos. Em geral, a modelagem da rede complexa é feita a partir dos parágrafos onde cada um representa um nó e a aresta define a afinidade destas sentenças. Porém, no ECC a proposta é modelar os textos baseado na construção linguística onde cada nó é uma palavra e as arestas indicam o sequenciamento de palavras e sua frequência (peso) já que o objetivo é buscar as ideias que prevalecem em um conjunto de textos sem perder e nem contrair informação. Outra diferença entre o ECC e os sumarizadores é que as frases extraídas pelo ECC são representações das construções linguísticas mais comuns, ou seja, representam o conteúdo relevante nos textos, enquanto que o sumarizador apenas seleciona parágrafos e elimina outros para obter um resumo. Desta forma, o ECC é capaz de obter uma multiplicidade grande de informações em torno de um tema ou palavra-chave, uma vez que apresenta probabilisticamente as sequências de palavras mais comuns do universo de textos, garantindo diversidade de informação. Além disso, o ECC deve ser capaz de permitir uma busca ativa de temas do interesse do usuário e mostrar as principais ideias de uma comunidade sobre um assunto que não seja necessariamente aquele mais relevante. Apesar do ECC não fazer sumarização, as técnicas utilizadas na sumarização extrativa inspiraram o desenvolvimento da ferramenta.

Este capítulo encerra a fundamentação teórica desta dissertação. A seguir será apresentada a arquitetura proposta para o Extrator de Conhecimento Coletivo bem como detalhes de sua implementação.

Capítulo 4

Arquitetura do Extrator de Conhecimento Coletivo

O Extrator de Conhecimento Coletivo é um instrumento de *Knowledge Discovery in Databases* (KDD) voltado para a informação social, sendo parte de um projeto mais amplo cujo objetivo é desenvolver uma plataforma de democracia digital e inteligência coletiva segundo a ideia de "Ágora Virtual" definida por Lévy (1999).

A proposta do ECC é ser o núcleo de processamento de dados desta plataforma, onde todos os participantes de um coletivo poderão enviar pequenos textos sobre questões de interesse comum e receberão tudo aquilo que se destaca no montante de informação, tornando possível conhecer seus principais desejos e anseios. Os objetivos se aproximam aos de uma ferramenta de coleta de dados sociais, tal como é desenvolvida no projeto *All Our Ideas* exemplificado na fundamentação teórica desta dissertação porém com metodologias distintas.

O ECC, visto como um projeto amplo de democracia digital, fundamenta-se no modelo de democracia participativa. Seu objetivo é tornar possível a participação direta do cidadão na tomada de decisão dos negócios públicos, uma vez que a expressão política passa a ser genuinamente da própria população e não de figuras representativas. Desta forma, é um instrumento que permite uma aproximação do que Silva (2005) definiu como o quinto grau da democracia digital onde as esferas pública e política voltam a se fundir e o cidadão passa a ter um papel ativo sobre as decisões que impactam a sua coletividade.

Para que tecnicamente este projeto seja possível, os fundamentos científicos do ECC devem ser criteriosamente escolhidos. A teoria de redes complexas mostra-se uma poderosa ferramenta para o estudo de sistemas complexos cuja causa do comportamento emergente não é facilmente determinada. A hipótese deste estudo é que a linguagem natural é um sistema complexo modelado ao longo de anos de história e evolução onde o conhecimento é uma resposta coletiva e emergente que garante a adaptação e sobrevivência dos seus agentes (e de si próprio) ao seu contexto. Nesta visão, a própria democracia também pode ser interpretada como um sistema complexo cujo comportamento emergente, que podemos entender como a expressão genuína da maioria, reflete a capacidade de uma população em se adaptar a um contexto e evoluir por si própria.

Além de fundamentar-se na teoria de redes complexas, a arquitetura do ECC também foi inspirada nas técnicas de sumarização extrativa e mineração de dados, as quais permitiram definir uma metodologia de processamento da informação e extração de ideias. A particularidade do ECC em relação a estas ferramentas é que este foca no que há de comum em um conjunto de opiniões enfatizando as opiniões majoritárias.

A arquitetura aqui proposta faz parte exclusivamente deste instrumento de processamento da informação. Porém, é importante ressaltar que sua aplicabilidade está condicionada a um contexto mais amplo que abarca o desenvolvimento de uma plataforma virtual onde usuários poderão interagir tanto inserindo informação quanto recebendo aquilo que está sendo processado. O objetivo final, o qual vai além dos objetivos desta dissertação, é torná-la parte de um sistema de web-democracia que tenha uma estrutura aplicável a todos os agentes de uma comunidade.

4.1 Arquitetura do ECC

O ECC é uma ferramenta que se propõe a coletar as informações mais relevantes de um banco de dados formado por uma grande quantidade de pequenos textos. Seu objetivo é classificar as informações e selecionar as mais representativas extraindo palavras-chave (temas) e parágrafos. A Figura 4.1 ilustra este processo segundo a abordagem KDD.

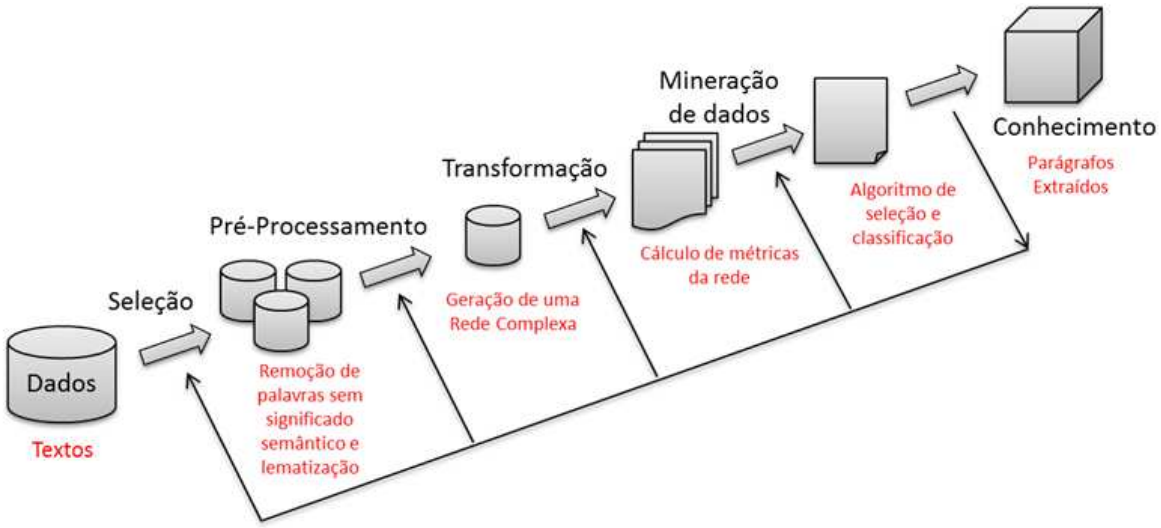

Figura 4.1: Etapas de processamento da informação no ECC segundo abordagem KDD (FAYYAD et al., 1996).

Para o cumprimento das etapas descritas na Figura 4.1, foi proposta uma arquitetura formada por quatro módulos distintos que interagem entre si e com fontes externas, conforme detalhado na Figura 4.2.

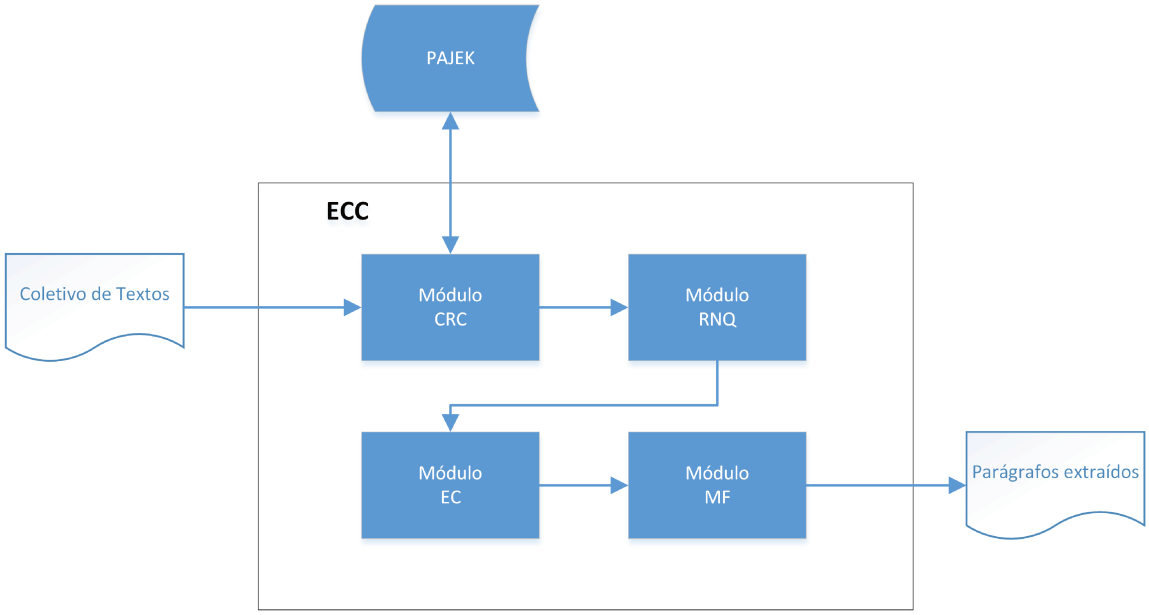

Figura 4.2: Arquitetura proposta para o Extrator de Conhecimento Coletivo. Os módulos recebem um coletivo de textos e entregam parágrafos. Um programa chamado PAJEK é utilizado para calcular as métricas de centralidade.

Como exemplo para ilustrar passo a passo o funcionamento do ECC, será utilizado o poema "No meio do Caminho" de Carlos Drummond de Andrade como banco de dados de entrada. Será apresentado como o texto está sendo manipulado, módulo a módulo, desde o pré-processamento até a extração do parágrafo final. Segue o poema:

> No meio do caminho tinha uma pedra
> Tinha uma pedra no meio do caminho
> Tinha uma pedra
> No meio do caminho tinha uma pedra.
>
> Nunca me esquecerei desse acontecimento
> na vida de minhas retinas tão fatigadas.
> Nunca me esquecerei que no meio do caminho
> tinha uma pedra
> tinha uma pedra no meio do caminho
> no meio do caminho tinha uma pedra.

4.1.1 Módulo CRC - Construtor de Rede Complexa

O Módulo Construtor de Rede Complexas (CRC) é o primeiro módulo da arquitetura ECC. Ele é responsável por receber o conjunto de textos e processá-lo gerando uma rede direcionada

baseada na co-ocorrência de palavras.

O conjunto de textos é composto por um documento único contendo todos os textos do banco de dados. Na primeira parte, as palavras contidas neste documento são tokenizadas (a ser definido adiante), rotuladas e lematizadas, processos executados por um "lematizador" desenvolvido pelo laboratório NILC (Núcleo Interinstitucional de Linguística Computacional) da USP cujo código-fonte está disponível para o uso (STEMMER, 2013). O objetivo destas etapas é simplificar e reduzir o texto melhorando seu processamento computacional sem perder informação semântica. Em seguida, depois de retirada as *stop-words*, o documento resultante é transformado em um grafo onde os vértices representam as palavras e as arestas indicam a co-ocorrência das mesmas no texto. O processo de construção da rede se dá a partir da leitura de cada parágrafo do documento, sobrepondo um ao outro através do incremento do valor das arestas (peso) ou da criação de arestas conforme a ocorrência de novas combinações de palavras.

Por fim, de posse da rede, duas métricas de centralidade são calculadas: grau e *betwenness*[1] e, caso o usuário deseje, o grafo é apresentado. Para o cálculo das métricas e visualização da rede foi utilizado o *software* PAJEK (BATAGELJ; MRVAR, 1998), um programa para análise e visualização de redes complexas. A figura 4.3 apresenta o fluxograma do Módulo CRC.

Descrição das etapas do fluxograma

1. Tokenização: consiste em decompor o texto em cada termo que o compõe (*tokens*). No geral, os termos estão separados por espaço em branco, quebras de linha e caracteres especiais (pontuações, entre outras).

2. Rotulação: uma vez reconhecidos os *tokens*, os mesmos são classificados de acordo com sua classe morfológica (artigo, adjetivo, substantivo, verbo, pronome, entre outras). Todos os tokens recebem um rótulo indicando sua natureza linguística.

3. Lematização: sabendo a classe morfológica de cada elemento do texto, o lematizador irá fazer a normalização morfológica de cada palavra reduzindo-a a sua forma canônica: passam-se os verbos para o infinitivo e os adjetivos e substantivos à forma masculina singular.

4. Eliminação das *stop-words*: são eliminadas as palavras de pouca relevância semântica para o processamento como os artigos, advérbios, preposições entre outras além das pontuações. Nesta etapa são gerados dois textos: Um contendo a pontuação final e outro sem a pontuação final. O primeiro será a base para construção da rede complexa. O segundo é necessário para o correto mapeamento em etapas posteriores. A figura 4.4 mostra o poema após o término desta etapa.

[1]Apesar do programa fazer o cálculo da métrica *betwenness*, a mesma não foi utilizada nos testes do ECC uma vez que em testes preliminares apresentou resultados semelhantes ao grau. No entanto, a implementação foi deixada caso haja necessidade de utilizá-la em trabalhos futuros.

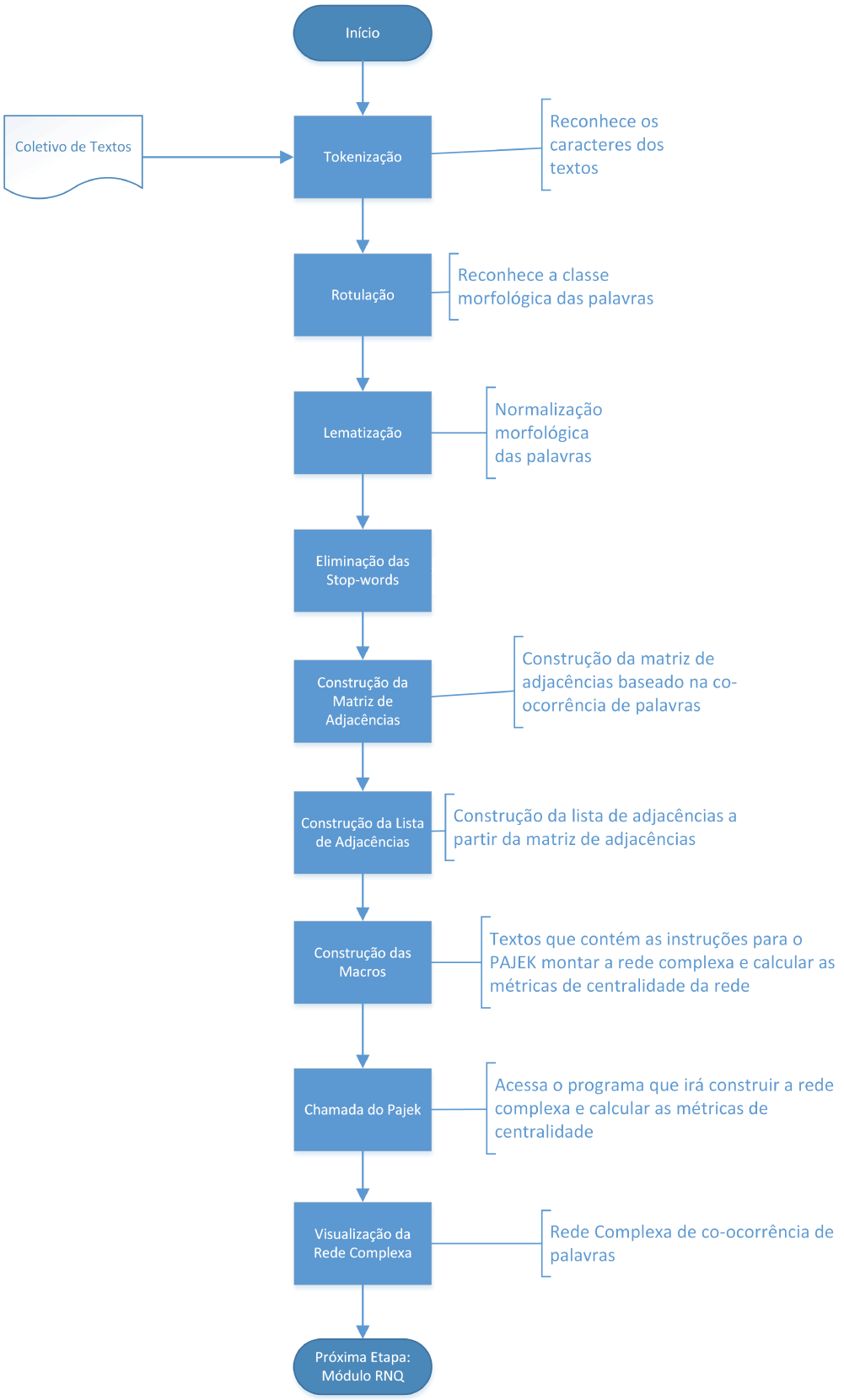

Figura 4.3: Fluxograma do Módulo CRC.

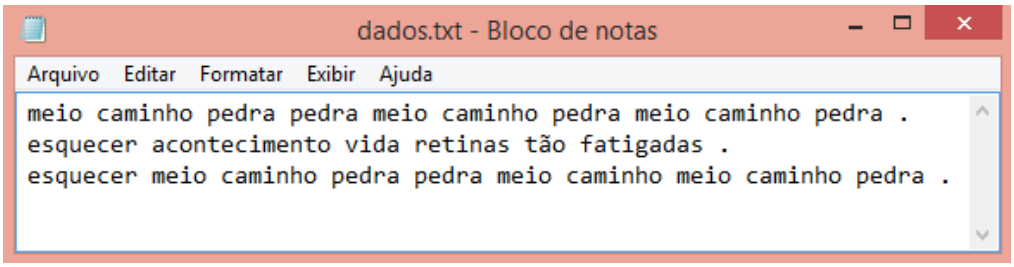

Figura 4.4: Poema lematizado e sem *stop-words*.

5. Construção da Matriz de Adjacências: uma vez pré-processado, o texto é transformado em uma rede complexa direcionada. As palavras são os nós da rede e as arestas indicam a co-ocorrência de palavras. A rede é construída baseada nos parágrafos do texto: para cada parágrafo é construída uma rede de co-ocorrência de palavras a qual se sobrepõe às redes dos outros parágrafos. A aresta contém um peso que indica quantas vezes aquela combinação de palavras ocorreu. A representação matemática de uma rede complexa se dá por uma matriz de adjacências: uma matriz quadrada cujos índices são os vértices da rede e os elementos representam a existência da ligação entre os dois índices. Por exemplo, o elemento a_{12} indica que o vértice 1 está ligado ao vértice 2. Vale lembrar que a rede é direcional, ou seja, a_{12} não é igual a a_{21}. O valor que o elemento assume representa quantas vezes aquela combinação de vértices (co-ocorrência de palavras) aconteceu, por exemplo, $a_{12} = 4$ representa que a ligação entre o vértice 1 e 2 ocorreu quatro vezes. A figura 4.5 (a) apresenta a matriz de adjacências do exemplo.

6. Construção da lista de adjacências: nesta parte do programa é gerada a lista de adjacências, a qual se trata de uma representação de rede complexa em formato de uma lista composta por três colunas: na primeira estão os vértices iniciais; na segunda estão os vértices finais; e na terceira a frequência que a combinação aconteceu. A figura 4.5 (b) apresenta a lista de adjacências do exemplo.

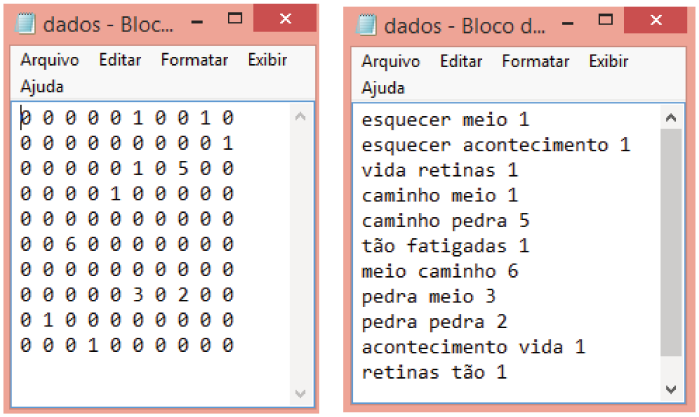

Figura 4.5: a) Matriz de adjacências e b) lista de adjacências referente ao poema lematizado e sem *stop-words*.

7. Construção das Macros: Para construção gráfica da rede complexa é utilizado um programa chamado PAJEK, o qual recebe dois arquivos: a lista de adjacências e a macro. A macro, que é criada nesta etapa, é um arquivo-texto que contém as instruções para o PAJEK construir a rede e calcular as métricas de centralidade.

8. Chamada do PAJEK: lista de adjacências e a macro são enviadas ao PAJEK.

9. Visualização da rede complexa: por fim, o PAJEK retorna a rede complexa e as métricas de centralidade calculadas e as apresenta ao usuário se assim desejar. A figura 4.6 ilustra a rede complexa do exemplo.

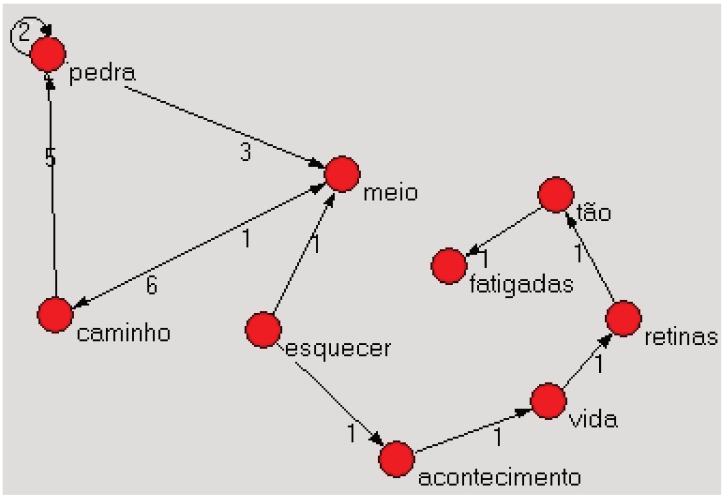

Figura 4.6: Rede complexa formada a partir do poema “No meio do caminho”.

4.1.2 Módulo RNQ - Ranqueador

O segundo módulo do ECC tem como objetivo gerar uma lista (*ranking*) formada por todos os vértices (palavras) da rede ordenados de forma decrescente segundo a métrica de centralidade escolhida pelo usuário. As métricas de centralidade são medidas de representatividade do nó e sua escolha deve se pautar nos atributos do sistema. Como a rede de co-ocorrência de palavras é um grafo orientado e com peso, foram implementadas para escolha do usuário as métricas grau e *betweenness* já que levam em consideração em seus cálculos o fato das arestas serem direcionadas e com peso[2]. Durante a execução do programa, o usuário deverá escolher qual métrica deseja para geração da Tabela *Ranking*. A figura 4.7 apresenta o fluxograma do Módulo RNQ.

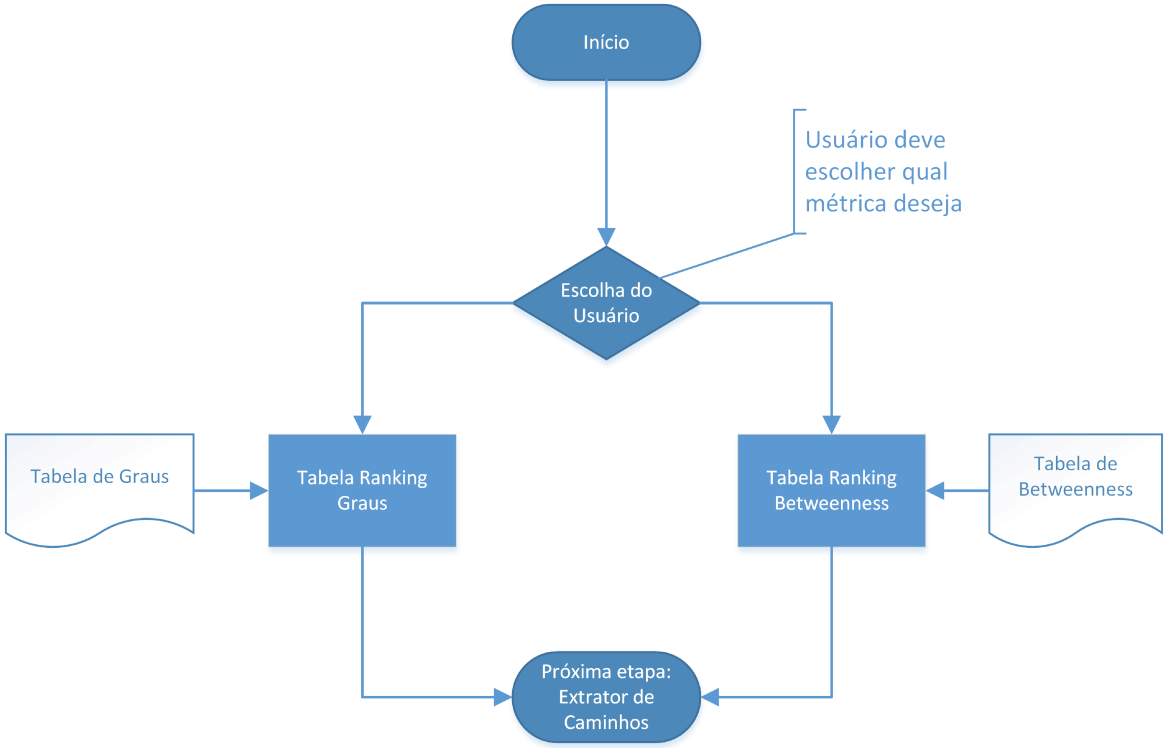

Figura 4.7: Fluxograma do Módulo RNQ.

Descrição das etapas do fluxograma

1. Escolha do usuário: o usuário deve escolher qual métrica será utilizada para pontuar os vértices da rede e gerar a Tabela *Ranking*: graus ou *betweenness*.

2. Apresentação da Tabela *Ranking*: É apresentada ao usuário a tabela correspondente à escolha feita na etapa anterior, com as palavras ordenadas do maior para o menor valor da centralidade escolhida. A figura 4.8 mostra a Tabela *Ranking* do exemplo segundo a métrica de centralidade grau.

[2]Apesar de ambas as métricas terem sido implementadas, optou-se por usar apenas a métrica Grau nos testes do ECC.

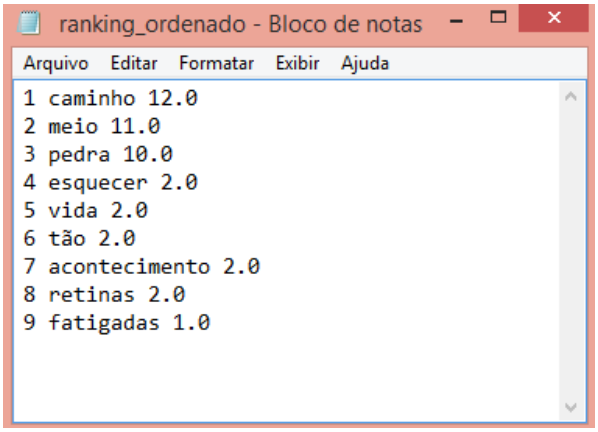

Figura 4.8: Tabela *Ranking* do poema usando a métrica de centralidade grau.

4.1.3 Módulo EC - Extrator de Caminhos

O objetivo desta etapa é buscar um conjunto de palavras, denominado proto-frase, aplicando um algoritmo que parta das palavras melhor classificadas na Tabela *Ranking* e "caminha" probabilisticamente pelo grafo segundo o peso de suas arestas. A cada passo do algoritmo sobre a rede uma palavra é selecionada para compor a proto-frase. As entradas deste módulo são as palavras escolhidas pelo usuário conforme a pontuação na Tabela *Ranking*, o número de palavras que irão compor a proto-frase (quantidade de passos do algoritmo) e a quantidade de proto-frases que deseja por palavra escolhida (número de vezes por palavra que o algoritmo de caminhos será executado). A saída é o conjunto de palavras que formam a proto-frase, a qual será entrada para o próximo módulo. A figura 4.9 apresenta o fluxograma do Módulo EC.

Descrição das etapas do fluxograma

1. Entradas do usuário: Inicialmente o usuário deve inserir quais palavras (vértices) deseja que sejam processadas, quantas palavras a proto-frase conterá e quantas proto-frases deseja para cada palavra.

2. Algoritmos de caminhos: Para cada palavra escolhida será aplicado um procedimento que caminha pela rede complexa e escolhe outras palavras segundo um sorteio probabilístico baseado no peso das arestas. Utiliza-se a matriz de adjacências para tal. Segue o algoritmo:

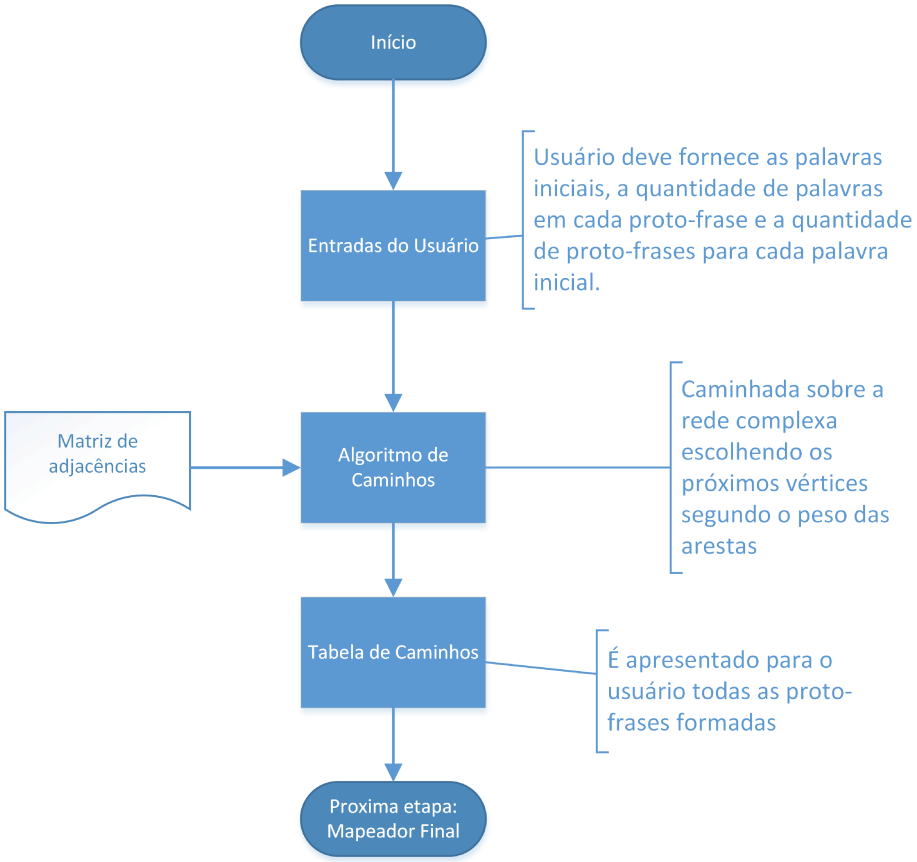

Figura 4.9: Fluxograma do Módulo EC.

Algoritmo 1: Algoritmo de Caminhos

Entrada: Palavra escolhida pelo usuário
Saída: Proto-frase
início
    Armazena a palavra na primeira posição da lista de caminhos;
    repita
        Some os pesos de todas as arestas que saem da última palavra armazenada na lista de caminhos;
        Sorteie probabilisticamente uma das arestas baseado em seus pesos;
        Identifique a palavra que a aresta vencedora aponta;
        Armazene a palavra na lista de caminhos;
    até *atingir a quantidade de palavras desejadas em cada proto-frase*;
fim

Lógica do Sorteio: Cada vértice possui arestas que chegam e saem dele e que contém um peso definido pela frequência da combinação entre duas palavras que aparece no conjunto de textos pré-processados. Um exemplo de uma rede complexa encontra-se na Figura 4.10.

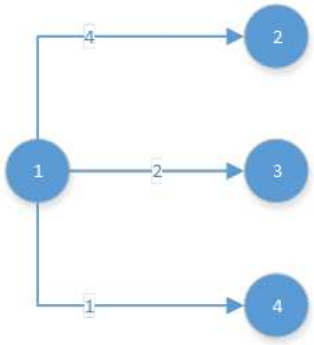

Figura 4.10: Exemplo de uma rede fictícia para ilustrar a lógica do sorteio.

Nota-se que esta é uma rede que contém quatro vértices e peso 4 para ligação $1-2$, peso 2 para ligação $1-3$ e peso 1 para ligação $1-4$. Supondo que o algoritmo está no vértice 1, a escolha do próximo vértice deve ser feita baseada no peso das arestas onde, quanto maior o peso maior é a probabilidade de ser escolhido o vértice indicado. Desta forma, o vértice 2 possuirá a maior probabilidade de ser escolhido seguido pelos vértices 3 e 4. Para que esta escolha seja feita segundo a distribuição de pesos, foi implementada uma função de geração de valor aleatório com distribuição não uniforme. Esta função lê o valor das arestas que saem do vértice e seleciona o próximo nó considerando a probabilidade ponderada pelos pesos.

3. Tabela de Caminhos: Completado os passos anteriores, é gerada uma lista com todas as proto-frases formadas, a qual é apresentada ao usuário, e passa-se para próxima etapa. A figura 4.11 apresenta dez proto-frases de 4 palavras cada extraídas do exemplo a partir da primeira palavra da Tabela *Ranking*.

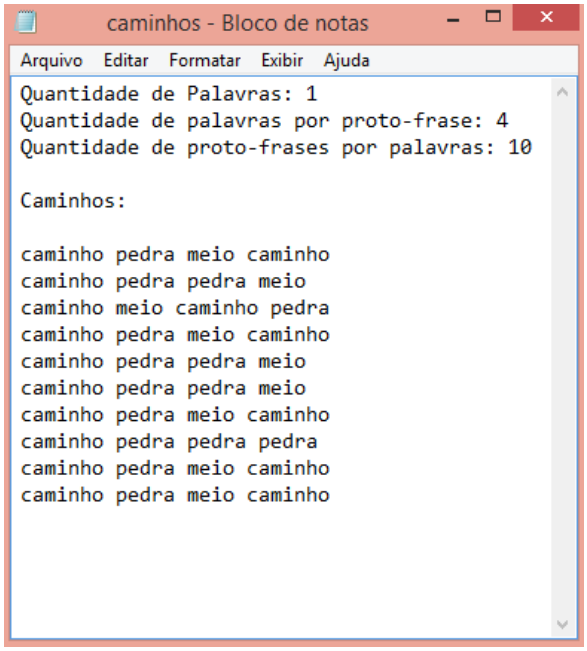

Figura 4.11: Proto-frases extraídas da rede formada pelo poema “No meio do caminho”.

4.1.4 Módulo MF - Mapeador Final

O último módulo do ECC tem como função extrair o parágrafo no documento de entrada que contém o trecho que melhor coincida com as palavras das proto-frases, tendo assim, os parágrafos que melhor representam o coletivo de ideias. Para cada proto-frase uma frase é extraída e apresenta-se ao usuário quantas palavras estão no parágrafo escolhido. A figura 4.12 ilustra o fluxograma do Módulo MF.

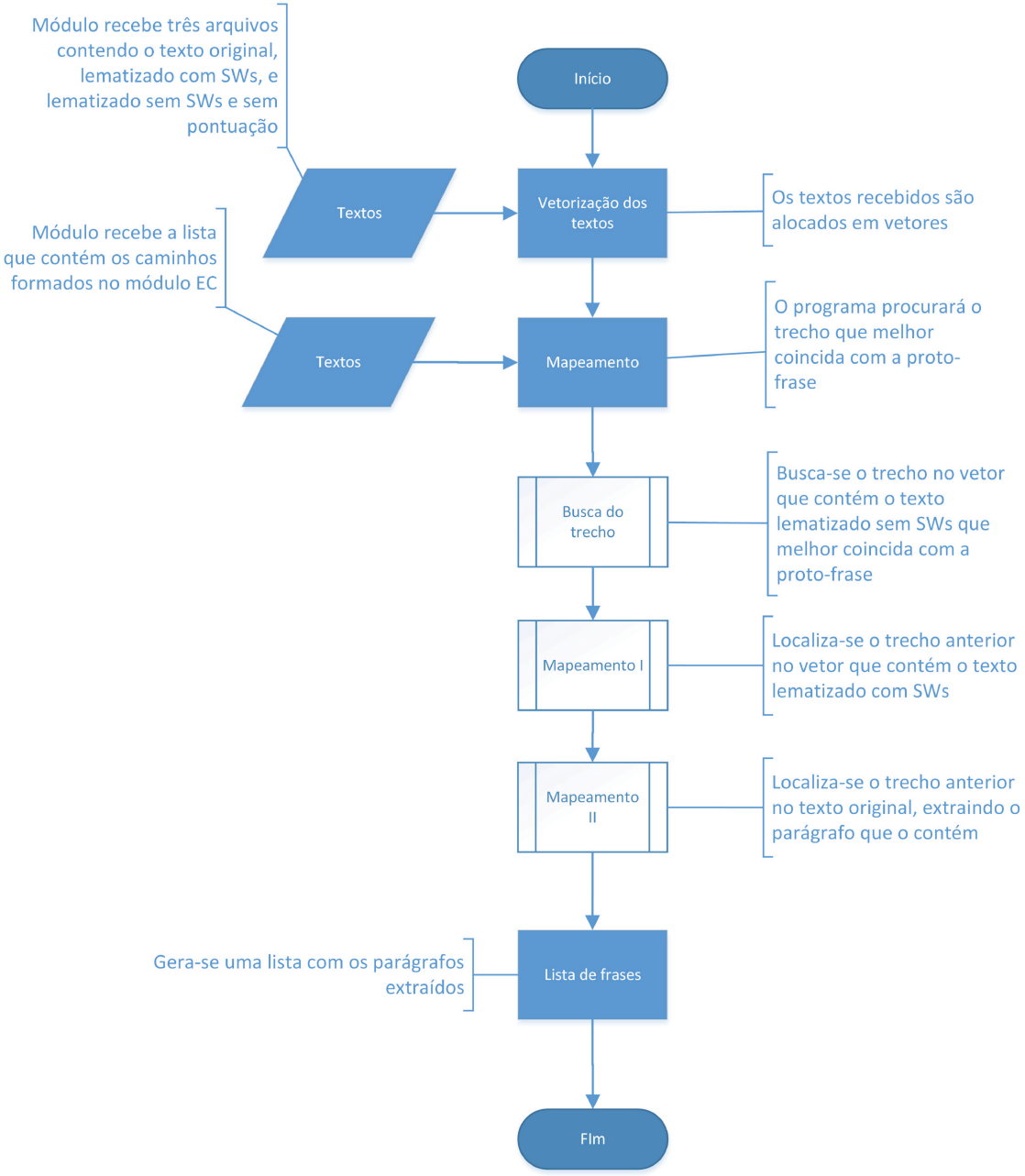

Figura 4.12: Fluxograma do Módulo MF.

Descrição das etapas do fluxograma

O Mapeador Final constitui-se de duas etapas: a vetorização dos textos e o mapeamento propriamente dito.

1. Vetorização dos textos: Nesta etapa, três textos são colocados em vetores: o documento original, o documento lematizado com *stop-words* (gerado pelo módulo CRC) e o documento lematizado sem *stop-words* e sem pontuação (também gerado pelo módulo CRC).

2. Mapeamento: o algoritmo de mapeamento consiste em duas partes: busca do trecho que melhor coincida com a proto-frase e o mapeamento deste trecho no documento original.

   (a) Busca do trecho: Inicialmente as palavras da proto-frase são colocadas em um vetor (vetor caminhos). É feita, então, uma varredura paralela, palavra por palavra, deste vetor com o vetor que contém o texto lematizado sem *stop-words* e sem pontuação. Em cada varredura é contado o número de palavras coincidentes na mesma posição relativa dos dois vetores. Para cada palavra coincidente, a combinação ganha um ponto. Assim, constrói-se um vetor (vetor *ranking*) cujo tamanho é o número de posições do vetor que contém o texto menos o número de palavras da proto-frase, onde, em cada posição, estará a pontuação referente à varredura entre os vetores. A figura 4.13 ilustra este procedimento utilizando como exemplo o poema, apresentando a situação dos vetores no passo 1, no passo 2 e no passo 6. Também apresenta a situação do vetor *ranking* ao final dos passos.

   Uma vez varrido todo o vetor texto, busca-se no vetor *ranking* a posição que contém o maior valor. No exemplo ilustrado, temos a posição 5 como aquela que contém o maior valor (4). Usa-se então esta posição como referência para encontrar o trecho no documento original.

   (b) Mapeamento I: O próximo passo é buscar o trecho no texto lematizado com as *stop-words*. O procedimento consiste em contar quantas palavras iguais (palavra cuja posição de referência aponta) há nas posições anteriores à posição de referência no vetor que contém o texto lematizado sem as *stop-words*. No exemplo, a palavra caminho está na posição de referência. Nas posições anteriores, a palavra caminho se repete uma vez. Assim, busca-se a posição em que a palavra caminho aparece pela segunda vez no texto lematizado com *stop-words* e armazena-a em uma variável de referência.

   (c) Mapeamento II: Como a quantidade de palavras do texto original é a mesma do texto lematizado com *stop-words*, a posição da palavra no vetor com o texto lematizado é a mesma no vetor com o texto original. Assim, tem-se a exata localização da palavra que é a referência do trecho procurado. A última etapa é extrair o parágrafo no qual se encontra a posição de referencia. Faz-se uma busca nas posições anteriores até o próximo ponto final ou início de um parágrafo e nas posições posteriores até encontrar um ponto final, armazenando as palavras em um vetor (vetor frase), o qual é apresentado ao usuário. A Figura 4.14 ilustra estes procedimentos.

Figura 4.13: Mapeamento 1 - Busca no trecho

3. Lista de frases: Por fim, o vetor frase é gravado em um arquivo .txt e é apresentado ao usuário em forma de lista de frases contendo a palavra escolhida, o parágrafo extraído, quantas palavras da proto-frase estão no parágrafo e o número de vezes que o parágrafo foi extraído. A figura 4.15 mostra os parágrafos extraídos do poema usado como exemplo para cada uma das 10 proto-frases. Nota-se que para as 10 proto-frases foi extraído o mesmo trecho do poema. Apesar da simplicidade do texto, o parágrafo extraído representa a ideia que mais se repete no poema.

Os resultados obtidos neste capítulo apenas ilustram didaticamente o funcionamento da arquitetura do Extrator de Conhecimento Coletivo. No próximo capítulo serão apresentados os resultados e a análise de um teste utilizando um banco de dados formado por um conjunto de

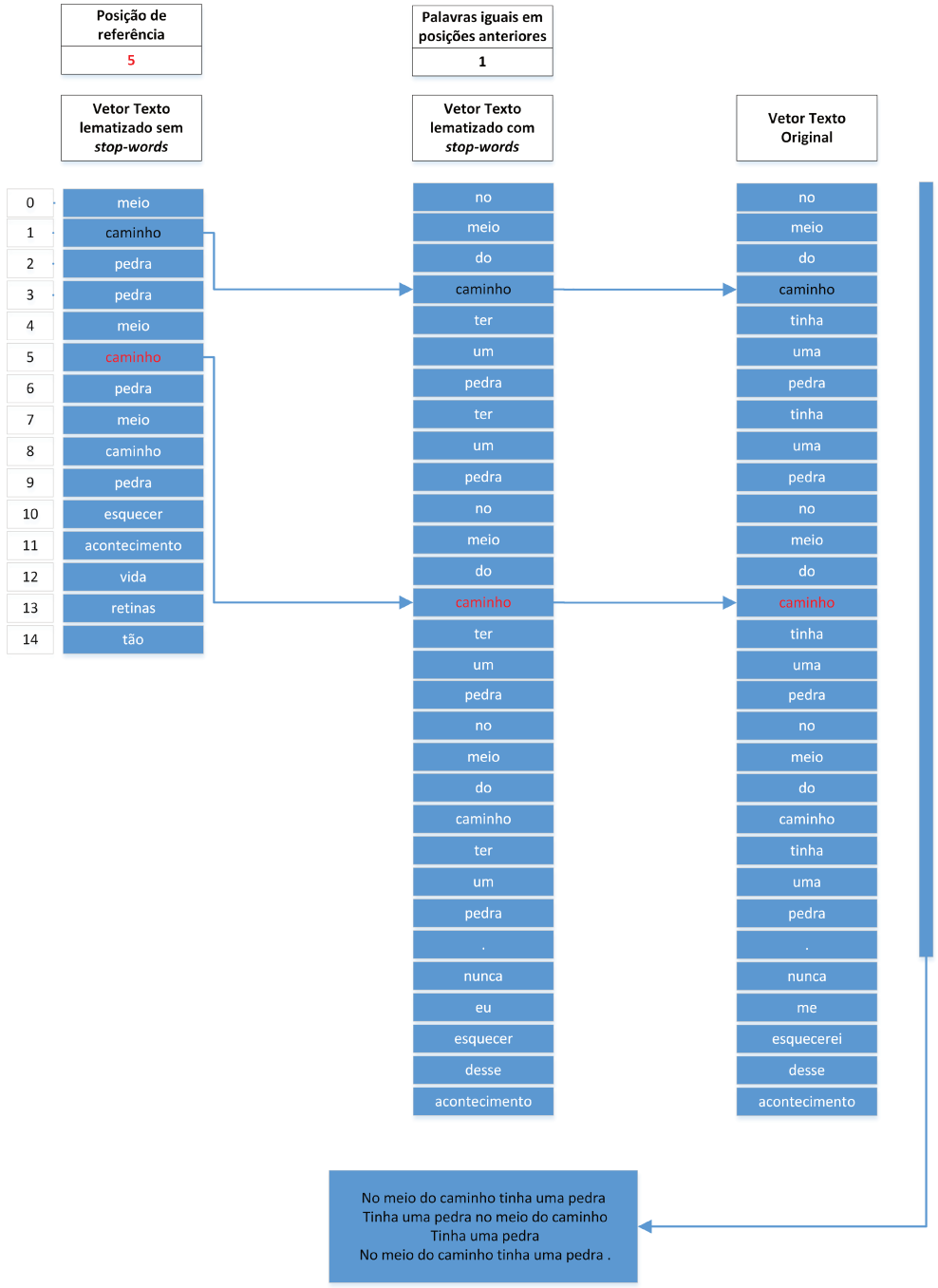

Figura 4.14: Mapeamento 2 - Busca no texto original e extração do parágrafo

textos.

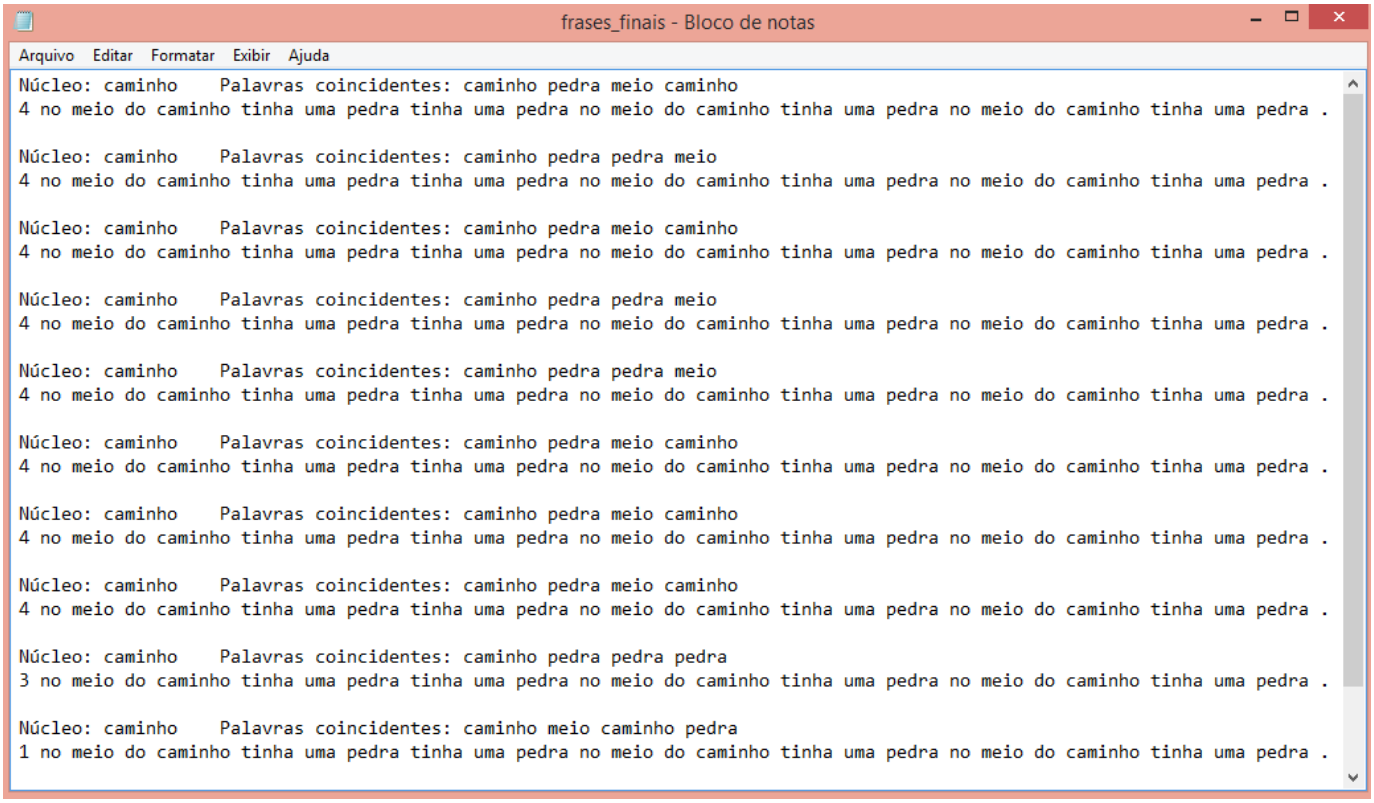

Figura 4.15: Parágrafos extraídos do poema após finalizadas as etapas de mapeamento.

Capítulo 5

Aplicando o ECC: conhecendo a opinião dos clientes de um Hotel

Com o objetivo de testar a arquitetura do ECC, foi desenvolvido um programa em linguagem Java o qual recebeu um banco de dados formado por 150 relatos de clientes sobre suas estadas em um hotel. Para apresentar os resultados e a discussão, o presente capítulo foi dividido em três partes: na primeira é feita uma caracterização do banco de dados, justificando a sua escolha e relatando o processo de preparo dos textos. Em seguida, o grafo formado é caracterizado e analisado conforme os modelos de redes complexas. Por fim, são apresentadas todas as etapas de processamento da informação desde a escolha das palavras e a formação das proto-frases até a extração dos parágrafos mais significativos.

5.1 Banco de Dados

Como um instrumento de coleta de dados sociais, a proposta do Extrator de Conhecimento Coletivo é processar pequenos textos escritos em linguagem natural pelos participantes de uma comunidade a respeito de assuntos da sua realidade e revelar os temas mais comentados e as sentenças que melhor representam a opinião coletiva a partir de uma busca ativa no banco de dados.

Para que os testes da metodologia apresentada no capítulo 4 ficassem o mais próximo possível dos seus objetivos, seria interessante que o banco de dados fosse constituído de relatos de membros de uma comunidade a respeito de assuntos comuns de sua realidade. Porém, devido a impossibilidade de adquirir estas narrativas durante a execução do presente projeto, a alternativa encontrada foi elaborar um banco de dados sobre um assunto suficientemente amplo que possibilitasse a busca de temas diversos e da opinião coletiva. Decidiu-se, então, conhecer como os clientes avaliam um hotel a partir de relatos públicos deixados no site *TripAdvisor*[1].

O *TripAdvisor* é um plataforma virtual voltada para promoção do turismo cujo conceito é fornecer informações com conteúdo gerado exclusivamente pelos usuários. Na descrição do *site*,

[1] http://www.tripadvisor.com

afirma ser o serviço mais popular e a maior comunidade de viagens do mundo com mais de 32 milhões de membros e mais de 100 milhões de comentários sobre hotéis, restaurantes, atrações e outros negócios relacionados a viagem. A escolha deste *site* está no fato de que, por ser uma comunidade virtual formada exclusivamente por opiniões de usuários, seu conteúdo pode ser classificado como "dados sociais", estando de acordo com tipo de informação que o ECC pretende explorar.

Dos assuntos disponíveis no *site*, foi escolhido o "Hotel" devido a generalidade do tema já que é possível explorar uma ampla gama de atributos, como por exemplo, avaliações sobre a qualidade do quarto, a localização do hotel, o café-da-manha, atendimento, serviço entre outros. Além disso, o próprio *site* disponibiliza para os membros que avaliam os hotéis uma pesquisa quantitativa onde devem pontuar de zero a cinco alguns itens pré-determinados. Essa pesquisa, apesar de simples, é um importante material para comparar e validar os resultados obtidos pelo ECC.

A escolha do hotel foi feita a partir da análise dos dados destas pesquisas. Buscou-se aquele em que as notas dadas aos itens pré-estabelecidos variassem a fim de que a mesma variação pudesse ser observada nos resultados obtidos pelo ECC. Dadas estas condições, foi selecionado um hotel com 279 avaliações escritas, sendo 180 feitas em português e o restante em outras línguas. Para compor o banco de dados foram utilizados 150 destes relatos. Um cuidado especial foi selecionar os comentários mais recentes já que comentários muito antigos podem referir-se a outros contextos fazendo as opiniões divergirem artificialmente. Todas as narrativas selecionadas foram publicadas no *site* entre janeiro e junho de 2014.

5.1.1 Etapa 1: Preparo do banco de dados

Os 150 relatos foram retirados do *site* na íntegra e alocados em um único documento. Para uma melhor eficiência no processamento do texto, foi feita uma revisão ortográfica do conteúdo, corrigindo palavras escritas erroneamente e substituindo numerais e abreviações pelas suas formas extensas. Esta etapa poderá ser incorporada futuramente na arquitetura do ECC e executada automaticamente. Por fim, todas as frases foram colocadas sequencialmente formando um único parágrafo. O documento final foi salvo com a extensão *.txt e inserido como entrada para o Extrator de Conhecimento Coletivo.

5.1.2 Etapa 2: Pré-processamento

Uma vez inserido o banco de dados, o programa implementado foi executado. O primeiro módulo (módulo CRC) preparou o documento para gerar uma rede conforme a metodologia descrita na seção 4.1.1. A Tabela 5.1 compara o banco de dados antes e depois do pré-processamento segundo a quantidade de palavras, de sentenças e a média de palavras por sentença. Esta última variável é uma importante referência para definir posteriormente o tamanho das proto-frases, uma vez que devem ser suficientemente grandes para mapear corretamente a sentença no *corpus* original porém não excessivamente maior do que as frases no documento pré-processado.

Tabela 5.1: Características do banco de dados antes e depois do pré-processamento.

Documento	**Palavras**	**Sentenças**	**Palavras/Sentença**
Antes do pré-processamento	9749	896	10.88
Após o pré-processamento	5256	896	5.87

5.2 Análise do grafo

Após o pré-processamento, o banco de dados foi mapeado como lista e matriz de adjacências as quais foram passadas para o programa PAJEK. Foram obtidas as seguintes métricas de centralidade para cada nó i e suas respectivas médias: Grau K_i e $\langle k \rangle$; Coeficiente de agrupamento C_i e $\langle C \rangle$; Comprimento do menor caminho L_i e $\langle l \rangle$. Também foram calculados o Diâmetro da rede D, o número de nós N e o número de arestas M. A Tabela 5.2 apresenta as propriedades indicadas e a Figura 5.1 a rede obtida.

Tabela 5.2: Propriedades da rede

Propriedade	**Símbolo**	**Valor**
Nós	N	1379
Arestas	M	3389
Diâmetro	D	16
Grau médio	$\langle k \rangle$	3.162
Coeficiente de agrupamento médio	$\langle C \rangle$	0.049
Caminho mínimo médio	$\langle l \rangle$	4.987

A partir das métricas apresentadas na Tabela 5.2, é possível caracterizar o grafo segundo os modelos de redes complexas descritos no capitulo 2.2.5. Porém, vale ressaltar que conhecer as métricas de centralidade não é condição suficiente para se determinar a estabilidade do modelo, uma vez que, para isto, faz-se necessário conhecer sua dinâmica de crescimento. Como a rede deste projeto representa um sistema invariante no tempo, não é possível afirmar que o modelo correspondente à sua topologia se manterá ao se inserir ou retirar dados da rede, porém uma vez caracterizada, é possível concluir a respeito de fenômenos como o efeito mundo pequeno ou o comportamento livre de escala. Para o propósito deste projeto, reconhecer a presença destes fenômenos é prioritário em relação a conhecer sua estabilidade, uma vez que busca-se determinar onde a informação se concentra e como navega pela rede.

Determinar um modelo de rede baseado na sua topologia não é uma tarefa trivial e, mais do que a análise direta das métricas de centralidade, envolve comparações com os valores esperados segundo especificações de cada modelo de rede. Conforme ilustrado na sessão 2.2.5, em redes regulares, os valores de $\langle l \rangle$ e $\langle C \rangle$ tendem a ser elevados, com o $\langle C \rangle$ próximo a 0.75 para redes muito acopladas. Além disso, uma rede completamente acoplada apresenta $\frac{N(N-1)}{2}$ arestas enquanto que em redes esparsas este número é na ordem de N ou N^2 . Ao analisar a Tabela 5.2, verifica-se que o coeficiente de agrupamento médio apresenta um valor 15 vezes menor do que os 0.75 esperados nas redes regulares. Do mesmo modo, caso o grafo exibisse propriedades

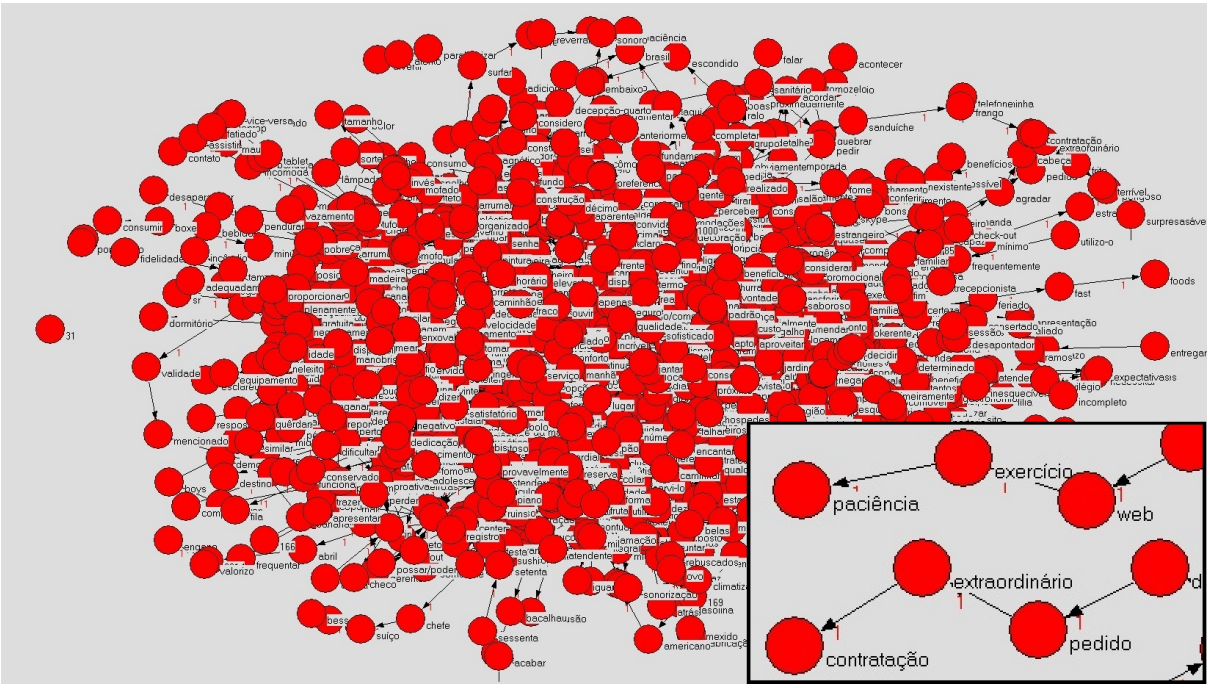

Figura 5.1: Rede de co-ocorrência de palavras obtida partir do banco de dados pré-processado. No detalhe, é mostrada a rede ampliada.

típicas das redes regulares, o número de arestas estaria na ordem de 950000, enquanto que o valor de M encontrado é de apenas 3389 arestas, próximo a ordem N de redes esparsas. Estes dados descartam a hipótese da rede obtida ser uma rede regular.

Em redes aleatórias do modelo de Erdős-Rényi, conforme sessão 2.2.5, é esperado um valor baixo para $\langle C \rangle$ e $\langle l \rangle$ e uma distribuição de graus que se aproxima de uma distribuição de Poisson. Neste modelo, o caminho mínimo médio encontra-se em torno de $\langle l \rangle \approx \frac{lnN}{ln\langle k \rangle}$. Substituindo N e $\langle k \rangle$ segundo valores da Tabela 5.2, $\langle C \rangle$ estaria próximo a 6.28. Porém, o valor real obtido de $\langle C \rangle$ foi 4.98, $20,1\%$ menor do que o esperado. O coeficiente de agrupamento médio de redes Erdős-Rényi encontra-se em torno de $\langle C \rangle \approx \frac{\langle k \rangle}{N}$, o qual, substituindo novamente pelos valores da Tabela 5.2, equivale a 0.00229, um valor 21 vezes menor do que o 0.049 observado. Em relação à distribuição dos graus da rede, a Figura 5.2 apresenta o gráfico obtido segundo os valores de k_i encontrados. É possível observar que o comportamento apresentado assemelha-se ao que é esperado para lei da potência com uma longa cauda e não como uma distribuição poissoniana característica de uma rede do modelo de Erdős-Rényi.

Para determinar se o grafo comporta-se como uma rede mundo pequeno, as métricas de centralidade devem ser comparadas aos valores que são esperados para as redes regulares e aleatórias já que é um modelo intermediário entre os dois. Um método para inferir se a rede apresenta características do modelo mundo pequeno é comparando os valores encontrados para $\langle l \rangle$ e $\langle C \rangle$ com os valores esperados para uma rede aleatória de mesmo grau médio e mesmo número de nós. A relação entre $\langle C \rangle$ calculado e o $\langle C_a \rangle$ de uma rede aleatória, onde $\langle C_a \rangle = \frac{\langle k \rangle}{N}$, deve seguir: $\frac{\langle C \rangle}{\langle C_a \rangle} \gg 1$. Substituindo os valores, obtêm-se $\frac{\langle C \rangle}{\langle C_a \rangle} \cong 21,37$. Do mesmo modo, a relação entre $\langle l \rangle$ calculado e $\langle l_a \rangle$ de uma rede aleatória, onde $\langle l_a \rangle = \frac{lnN}{ln\langle k \rangle}$ deve seguir: $\frac{\langle l \rangle}{\langle l_a \rangle} \approx 1$.

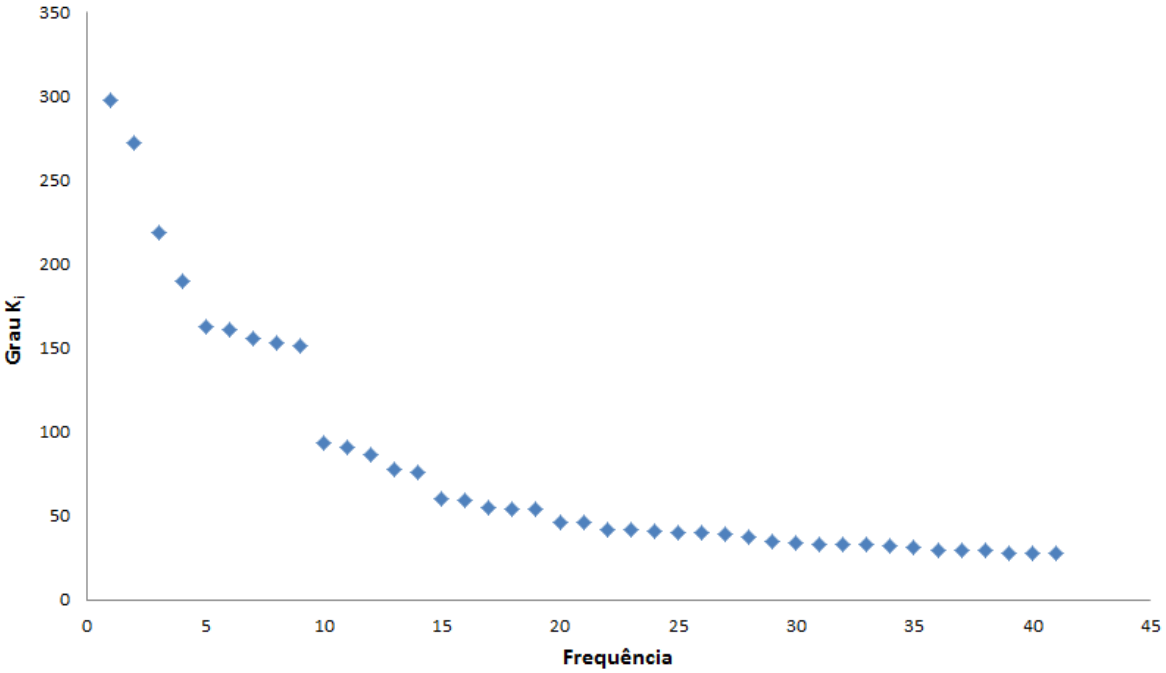

Figura 5.2: Frequência de distribuição de graus dos nós da rede.

Substituindo segundo os valores da tabela 5.2, encontra-se $\frac{\langle l \rangle}{\langle l_a \rangle} \cong 0.8$. Estes resultados indicam a possibilidade da rede estar mais próximo ao modelo mundo pequeno, fato este reforçado pelo valor de $\langle l \rangle$ (4.98) que claramente demonstra o efeito mundo pequeno já que a distância mínima média entre os nós é baixa e próxima dos valores encontrados nas redes mundo pequeno descritas na sessão 2.2.5.

Por fim, pode-se investigar se o grafo apresenta características de redes livres de escala, conforme descrito na sessão 2.2.5. Para isto, deve-se analisar a frequência de distribuição de graus dos vértices e determinar se o seu comportamento segue a lei de potência, com γ variando entre 1 e 3 (BARABÁSI; ALBERT; JEONG, 1999). Em uma análise da figura 5.2 é possível observar que poucos nós estão muito conectados e muitos nós pouco conectados arrastando-se em uma cauda longa, comportamento típico da lei de potência. Segundo demonstrado na sessão 2.2.5, a lei de potência pode ser descrita na seguinte forma

$$P(k) = C.k^{-\gamma} \tag{5.1}$$

onde C e γ são constantes e C representa uma aleatoriedade tipicamente na ordem de e^c (NEWMAN, 2010). Em um gráfico $log - log$, a lei de potência pode ser expressa da mesma forma de uma equação

$$logP(k) = -\gamma.log(k) + logC. \tag{5.2}$$

Desta forma, γ é facilmente determinado através do coeficiente angular da reta que melhor se aproxima aos pontos. A Figura 5.3 apresenta a distribuição de graus em escala logarítmica bem como a equação da reta melhor aproxima os pontos usando o método de regressão linear.

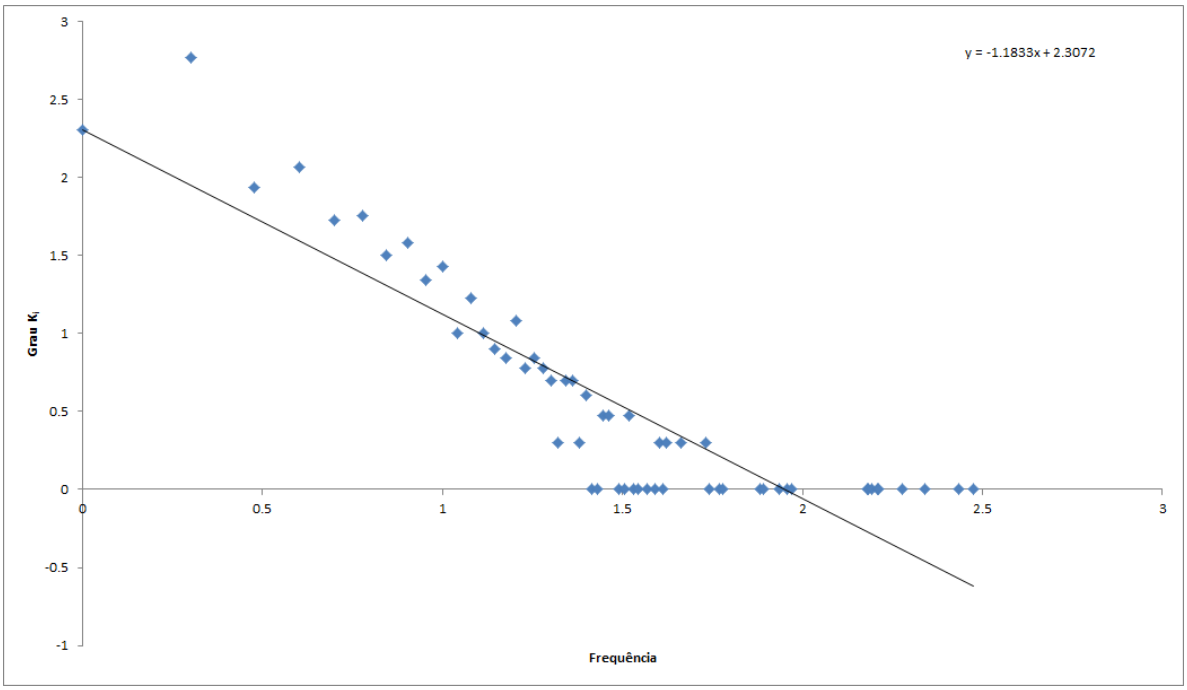

Figura 5.3: Frequência de distribuição de graus dos nós da rede em escala logarítmica e reta e equação da reta que melhor aproxima os pontos.

Como é possível observar através da equação da reta na Figura 5.3, $\gamma = 1.18$, valor dentro da faixa esperada para uma rede que apresenta comportamento livre de escala.

Analisando os resultados da caracterização do grafo, a rede obtida a partir da co-ocorrência de palavras é uma rede que apresenta tanto o fenômeno mundo pequeno quanto o comportamento livre de escala, corroborando com os resultados de pesquisas que apontam este tipo de rede como uma rede mundo pequeno. Estas características não apenas permitem classificar o conjunto de textos como um sistema complexo como também garantem uma organização da informação na topologia da rede através de *hubs* e caminhos curtos. Tais resultados implicam em duas condições:

1. A presença do efeito mundo pequeno indica que com poucos passos pode-se alcançar quase todos os pontos da rede, tornando-a facilmente navegável na busca da informação relevante.

2. O comportamento livre de escala indica a presença de *hubs* na rede, pontos centrais que concentram informação e serão o ponto de partida na busca da informação coletiva.

5.3 Processamento e extração da informação coletiva

Uma vez caracterizado o grafo através de um modelo de redes complexas e concluído que o mesmo comporta-se segundo fenômenos esperados para sistemas complexos, é possível iniciar o tratamento da informação com o intuito de extrair seu conteúdo mais relevante, conforme a metodologia apresentada no capítulo 4. A primeira etapa consiste na análise da Tabela Ranking elencando os temas mais importantes . Em seguida é aplicado o Algoritmo de Caminhos e retiradas as proto-frases. Por fim, é realizado o mapeamento no conjunto de textos original e extraídas as frases que representam o conhecimento coletivo. Estas etapas serão detalhadamente descritas a seguir.

5.3.1 Tabela Ranking

A Tabela Ranking apresenta os nós em ordem decrescente segundo seu valor de grau. Como uma métrica de centralidade, quanto maior o grau de um nó mais central ele está na topologia da rede, logo, mais informação ele carrega. A Tabela 5.3 mostra a Tabela Ranking obtida para os 25 nós de maior grau e a Figura 5.4 apresenta todos os nós da rede segundo seu valor de grau.

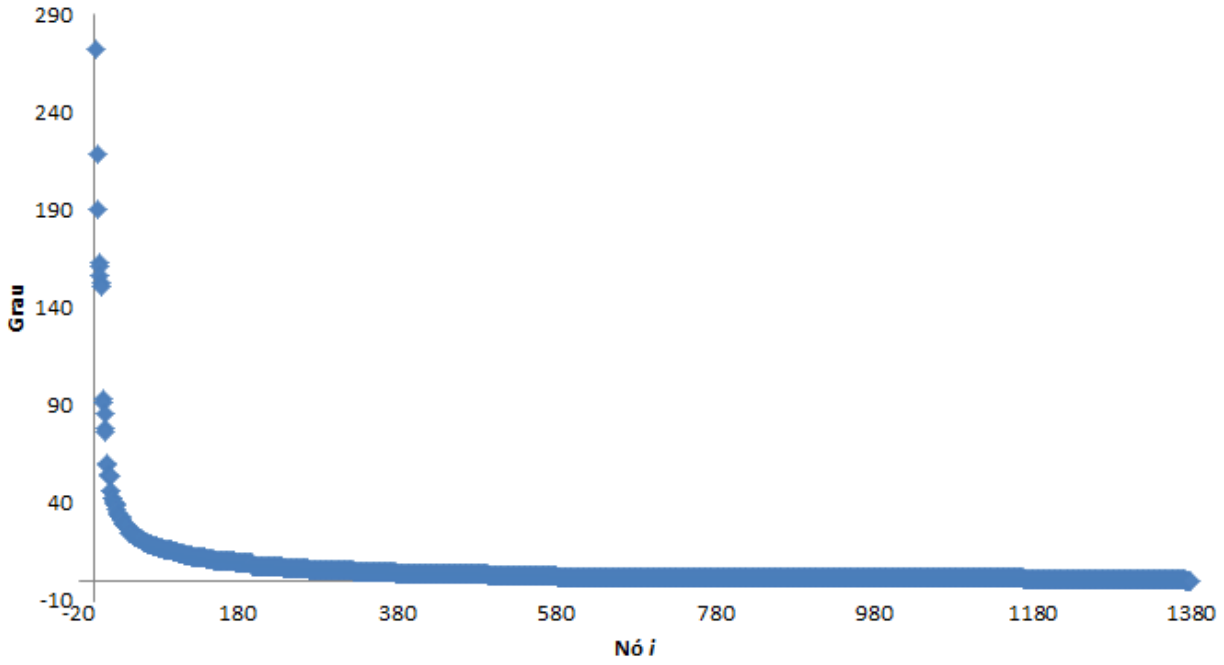

Figura 5.4: Valor do grau para cada nó da rede apresentado em ordem decrescente conforme Tabela Ranking.

Da análise da Tabela 5.3 e da Figura 5.4, observa-se que a maioria dos nós possuem um valor de grau próximo a zero conforme já verificado na distribuição de graus. Porém, os nós de maior grau concentram valores próximos que se agrupam em pequenos conjuntos ou nitidamente se isolam dos demais. Para verificar esta constatação, a Figura 5.5 apresenta as distâncias relativas em graus entre os 31 nós de maior grau.

Tabela 5.3: Tabela Ranking dos 25 nós de maior grau.

Colocação	**Palavra**	**Grau**
1	hotel	298
2	bom	272
3	quarto	219
4	bem	190
5	localização	163
6	café	161
7	paulista	156
8	manhã	153
9	excelente	151
10	restaurante	93
11	avenida	91
12	próximo	86
13	atendimento	78
14	confortável	76
15	serviço	60
16	localizar	59
17	cama	55
18	ótimo	54
19	opção	54
20	metrô	46
21	banheiro	46
22	limpo	42
23	shopping	42
24	negócio	41
25	paulo	40

Analisando a Figura 5.5, nota-se um isolamento dos nós 1, 2, 3 e 4, com destaque a distância entre os nós 2 e 3. Já a partir do quinto nó nota-se a presença de agrupamentos (nós com distâncias próximas de zero). Um primeiro conjunto é observado entre os nós 5 e 8 e um segundo conjunto entre os nós 10 e 13. A partir do nó 15 é mantida uma distância média próxima de zero.

Em relação ao conteúdo dos vértices, o nó de maior grau refere-se a palavra “hotel”, o tema central do banco de dados. Já o segundo nó refere-se ao adjetivo “bom”, sendo possível inferir que o hotel seja o tema representativo do conteúdo dos textos e o adjetivo “bom” a avaliação mais citada. Este último dado corrobora a pesquisa quantitativa mostrada na Figura 5.6, a qual indica que 76% dos avaliadores consideram o hotel excelente/bom. Já o terceiro nó refere-se a um atributo do hotel, no caso o quarto, seguido pelo advérbio “bem”. Em seguida vem um conjunto de nós com valores muito próximos que se referem basicamente a classes e qualidades do hotel: localização, café, paulista, manhã e excelente. A Tabela 5.4 classifica os 15 nós de

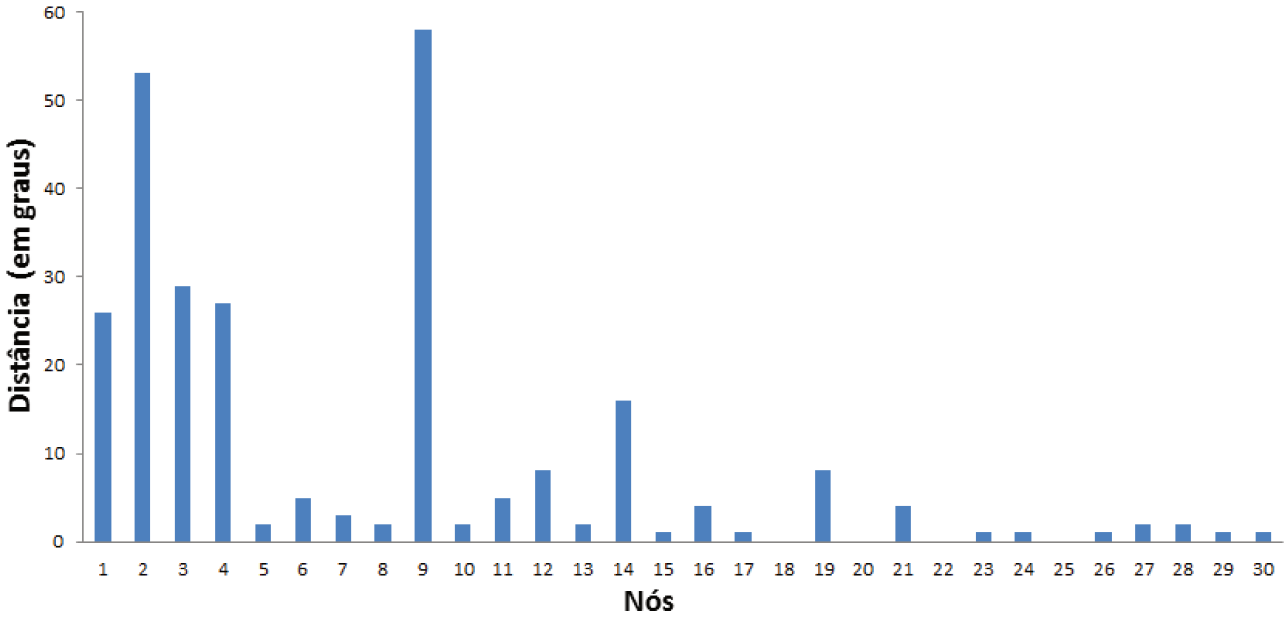

Figura 5.5: Distâncias relativas entre os 31 nós de maior grau. Os valores do eixo x representam os pares de nós segundo colocação da Tabela Ranking: O 1 é a distância entre os nós 1 e 2, o 2 entre os nós 2 e 3 e assim por diante.

maior grau segundo o conteúdo ao qual se referem, sendo possível inferir que Hotel é o tema geral e que os principais atributos comentados pelos clientes estão o quarto, a localização, alimentação (café, restaurante) e o serviço (atendimento). Os outros itens dizem respeito a percepções dos avaliadores (bom, excelente, confortável)..

Tabela 5.4: Classificação proposta para os 15 nós de maior grau.

Tema	**Grau**	**Atributos**	**Grau**	**Qualidade**	**Grau**
hotel	298	quarto	219	bom	219
		localização	163	bem	190
		café	161	excelente	151
		restaurante	93	confortável	76
		atendimento	78		

Estes dados podem ser confrontados com os resultados da pesquisa apresentada na Figura 5.6. O campo “Resumo das pontuações” apresenta os seis itens mais avaliados pelos clientes em um universo de 10 itens, os quais correspondem aos atributos da Tabela 5.4: quarto (qualidade do sono e quarto), localização e serviço (atendimento e limpeza). Já em relação ao campo “Pontuação dos viajantes”, 51% dos clientes consideram o hotel muito bom enquanto que 25% o classificaram como excelente. Este resultado acompanha aquele apresentado na Tabela 5.4 onde o grau de “bom” é 45% maior do o grau de “excelente”, indicando a predominância do primeiro adjetivo sobre o segundo.

Figura 5.6: Pesquisa quantitativa extraída da página do hotel no site *TripAdvisor*

A Tabela Ranking permite conhecer os pontos mais relevantes da topologia da rede e inferir uma classificação para os nós buscando grupos de atributos, temas mais comentados e avaliações generalistas. A partir dela são selecionados os nós para que seja possível extrair as frases que representam o conhecimento coletivo. Para prosseguir com os testes, foram selecionados os atributos presentes na tabela 5.4: quarto, localização, café, restaurante e serviço; para serem o ponto de partida do Algoritmo de Caminhos para geração das proto-frases.

5.3.2 Algoritmo de Caminhos

Escolhidos os vértices que servirão como ponto de partida para o Algoritmo de Caminhos, duas variáveis devem ser definidas: a quantidade de palavras de cada proto-frase e a quantidade de proto-frases que será retirada para cada palavra.

Quantidade de palavras por proto-frase

Definir o número de palavras da proto-frase é uma tarefa importante na extração das frases mais significativas. Uma proto-frase muito pequena pode levar a uma grande quantidade de frases, não havendo convergência da informação. Por outro lado, uma proto-frase muito grande pode dificultar a localização de todas as palavras, ou ao menos a maioria, em uma única sentença.

Uma medida interessante para se determinar o tamanho das proto-frases é o comprimento médio, em palavras, das frases do banco de dados após a lematização e retirada das *stop-words*. Através da Tabela 5.2, é possível notar que, após processado, o texto possui, em média, 5.8 palavras por frase. A metodologia adotada será definir este valor como um limite superior para o tamanho da proto-frase, já que proto-frases maiores diminuem a probabilidade de encontrar todas suas palavras em uma única sentença. Para testar essa hipótese, foram realizados 10 testes variando o comprimento da proto-frase de 1 a 12 palavras e executando, em cada teste, o algoritmo de caminhos 50 vezes. O nó de partida será o referente à palavra "Hotel". A Tabela 5.5 apresenta os resultados do teste, mostrando quantas sentenças repetidas foram extraídas a partir das 50 proto-frases formadas de comprimento variando entre 1 a 12 palavras.

Para proto-frases de até 3 palavras, não foram extraídas sentenças repetidas, o que implica em proto-frases generalistas que não convergem para uma frase. Já em proto-frases com 6 ou mais palavras a quantidade de sentenças repetidas aumenta quase linearmente, indicando

Tabela 5.5: Resultado do teste para definir a metodologia de escolha do comprimento da proto-frase.

Comprimento da proto-frase	**Quantidade de proto-frases repetidas**
1	0
2	0
3	0
4	2
5	5
6	8
7	7
8	8
9	9
10	10
11	13
12	17

uma redundância de informação. Para garantir alguma diversidade sem redundância, o valor escolhido para o comprimento da proto-frase será de 5 palavras, limite inferior da média de palavras por sentença do texto lematizado e sem *stop-words*.

Quantidade de proto-frases por palavras

Outra variável importante a ser definida é a quantidade de proto-frases que serão geradas para cada palavra escolhida, ou seja, quantas vezes o Algoritmo de Caminhos será executado. Uma vez que o algoritmo caminha probabilisticamente pelos nós, é importante executá-lo diversas vezes para que as proto-frases convirjam para o(s) trecho(s) mais significativos sem perder a diversidade da informação e nem capturar muitos trechos que não mais expressam o conhecimento coletivo. Como a decisão de cada passo está diretamente relacionadas ao peso das arestas, uma possibilidade de determinar este parâmetro é utilizando a métrica graus de saída (*outdegree*), ou seja, a quantidade de arestas que saem de cada nó.

A Tabela 5.6 apresenta os 15 maiores graus de saída. Para testar a hipótese descrita, foi realizado um teste partindo da palavra "Hotel" e executando o algoritmo de caminho 102 vezes gerando proto-frases de 5 palavras, o qual resultou em 102 proto-frases extraídas onde 99 são distintas e apenas 3 se repetem duas vezes. Este resultado reflete a adequadamente a diversidade que deve gerar o tema Hotel com baixa redundância de informação.

Em resumo, as metodologias adotadas para o cálculo dos parâmetros do Algoritmo de Caminhos serão: o comprimento da proto-frase será o limite inferior do tamanho médio de uma frase do documento lematizado e sem *stop-words*;e o número de vezes que o algoritmo de caminhos será executado para cada palavra selecionada será o valor do grau de saída do nó inicial. A

Tabela 5.6: Palavras com os 15 maiores graus de saída.

Palavra	**Grau de Saída**
hotel	102
bom	78
quarto	73
bem	47
localização	45
paulista	42
manhã	39
restaurante	35
excelente	32
atendimento	30
confortável	27
banheiro	21
localizar	20
serviço	18
cama	18

próxima etapa consiste em mapear as proto-frases no texto original e extrair os parágrafos.

5.3.3 Mapeamento

Selecionadas as palavras mais significativas utilizando-se da Tabela Ranking e definidos os parâmetros do Algoritmo de Caminhos, a última etapa consiste em extrair as sentenças mais significavas através do mapeamento da proto-frase no banco de dados. O mapeamento consiste em buscar nos textos originais o trecho equivalente à série de palavras da proto-frase. Porém, é comum que durante sua execução nem todas as palavras da proto-frase estejam presentes na mesma sentença, levantando à necessidade de fazer uma seleção das frases extraídas segundo os termos coincidentes. Para que as frases representem o mapeamento da melhor forma possível, foi definido que seriam consideradas apenas aquelas que possuírem todas as palavras da proto-frases. Caso não haja nenhuma sentença nesta condição, serão considerada(s) a(s) frase(s) que contém a maior quantidade de palavras da proto-frase. Assim, ao final, para cada conjunto de proto-frases apenas uma porcentagem delas gerará, de fato, uma sentença extraída. A seguir serão apresentados os resultados e análise de cada etapa do mapeamento para cada palavra escolhida segundo definido na sessão 4.1.4.

Quarto

O atributo de maior grau, segundo a Tabela 5.4, é a palavra quarto, a qual será ponto de partida para executar o algoritmo de caminhos e o mapeamento. A Tabela 5.7 apresenta os

paramêtros utilizados para o teste, a Tabela 5.8 as 73 proto-frases formadas e a Tabela 5.9 o resultado final do mapeamento, apresentando as frases extraídas, a quantidade de palavras da proto-frase que contém e quantas vezes foi mapeada pelas proto-frases.

Tabela 5.7: Parâmetros para execução do algoritmo de caminhos partindo da palavra "quarto".

Palavra	quarto
comprimento da proto-frase	5
número de proto-frases	73

Tabela 5.8: Proto-frases formadas pela execução do algoritmo de caminhos partindo da palavra "quarto".

Proto-frases
quarto alto cidade lugar melhor
quarto antigo fórmula quarto escuro
quarto apesar reforma estender dia
quarto baixo temporada
quarto barulhento noite sono ruim
quarto barulho durante café manhã
quarto bastante quarto lavanderia hotel
quarto bastante velho lâmpadas mau
quarto bem conservar café manha
quarto bem diversificado excelente atendimento
quarto bem espaçoso wi-fi bom
quarto bem localizar próximo metrô
quarto bem servido bom lobby
quarto bem sortido
quarto bioma cama extra
quarto bom coração alameda santos
quarto bom custo benefício qualidade
quarto bom oferecer quarto ruim
quarto bom oferecer serviço recepção
quarto bom prato cobrar balcão
quarto bom serviço café manhã
quarto café almoço achar banheira
quarto confortável bem localizar novo
quarto confortável convidar
quarto confortável convidar
quarto confortável ducha água vazar
quarto confortável funcional especialmente termo
quarto confortável oferecer café manha
quarto confortável roupa cama deixar
quarto confortável seguro subir andar
quarto confortável trabalhar quarto atendimento

quarto dar azar quarto variado
quarto deixar desejar limpeza manutenção
quarto descuidado precisar voltar buscar
quarto entretanto serviço arrumação hora
quarto entretanto serviço jantar maravilhoso
quarto entretanto serviço quarto excelente
quarto espaçoso bem localizar próximo
quarto espaçoso cama razoável sentir
quarto espaçoso decorados adequadamente
quarto espaçoso hotel realmente movimentar
quarto espaçoso limpeza impecável ar
quarto espaçoso limpeza quesito destacar
quarto espaçoso limpo bem localizar
quarto espaçoso wi-fi bom custo-benefício
quarto excelente opção paulo trabalho
quarto excelente qualidade decair desde
quarto frente caminhão poder/podar escolher
quarto gostar confortável espaçoso roupa
quarto gostar hotel ótimo localização
quarto ingerir parte comercial metrô
quarto ingerir parte cozinha então
quarto lado bom embora preço
quarto lado paulista restaurante variadas
quarto lavanderia hotel bem conservado
quarto lavanderia hotel situar apenas
quarto limpo bem satisfatório bem
quarto limpo café manhã ótimo
quarto limpo claro desejar demora
quarto limpo confortável comido ótima
quarto melhor hotel preço equivalente
quarto novo lobby melhor quarto
quarto parede perto aproveitar bastante
quarto possuir piscina atrativo
quarto recomer/recomendar ótima tarifa promocional
quarto reposição consumo
quarto ruim lento necessário recomer/recomendar
quarto sentir desinteressar faltar cortina
quarto silencioso bem servido bom
quarto silencioso localização imponência lobby
quarto simples hotel super prestativo
quarto vinte quatro hora manhã

Segundo a Tabela 5.9, o número máximo de palavras das proto-frases encontradas nas sentenças extraídas foi 4. Quatro frases foram mapeadas uma única vez, cujo conteúdo refere-se ao quarto. Analisando o conteúdo das sentenças, é possível concluir que a opinião geral a respeito

Tabela 5.9: Frases extraídas após mapeamento da palavra "quarto" indicando o número de palavras contidas na proto-frase e quantas vezes sua extração se repetiu.

N. de Palavras da proto-frase	**Repetições**	**Frase**
4	1	Possui serviço de quarto vinte e quatro horas, o atendimento dos funcionários é cordial.
4	1	Os quartos são bem espaçosos.
4	1	Limpo quarto espaçoso , com wi-fi.
4	1	Tem quartos espaçosos e decorados adequadamente.

deste atributo é positiva e ressalta características como o fato de "ser espaçoso", "limpo", "decorado" e "com wi-fi". Comparando estes resultados com a pesquisa da Figura 5.6, nota-se que o quarto possui uma boa avaliação (quatro e meio de cinco pontos) corroborando com os achados do ECC.

Localização

O segundo atributo de maior grau é a palavra localização, a qual será o novo ponto de partida para executar o algoritmo de caminhos e o mapeamento. A Tabela 5.10 apresenta os paramêtros utilizados para o teste, a Tabela 5.11 as 45 proto-frases geradas e a Tabela 5.12 o resultado final do mapeamento, apresentando as frases extraídas, a quantidade de palavras da proto-frase que contém e quantas vezes foi mapeada pelas proto-frases.

Tabela 5.10: Parâmetros para execução do algoritmo de caminhos partindo da palavra "localização".

Palavra	localização
comprimento da proto-frase	5
número de proto-frases	45

Tabela 5.11: Proto-frases formadas pela execução do algoritmo de caminhos partindo da palavra "localização".

Proto-frases
localização atendimento valer pena horário
localização bairro jardim pertinho shopping
localização bom bom excelente localização
localização bom café-da-manhã refeição cansado
localização bom custo benefício qualidade

localização bom estrutura bom cama
localização bom localização conforto dinheiro
localização bom serviço hotel espaçoso
localização bom tempo desde atendimento
localização conforto hotel três quadra
localização conseguir realmente lindo quarto
localização conseguir taxis local café
localização coração paulo quadra metrô
localização custo benefício qualidade pessoa
localização destaque excelente hotel bem
localização excelente café manhã bastante
localização excelente custo-benefício bem conferência
localização excelente localização ótima tarifa
localização excelente serviço despertador funcionar
localização excepcional adequado possível conseguir
localização hotel centro negócio perfeito
localização jardim pertinho paulista restaurante
localização nada aparente reclamar reposição
localização nada aparente reclamar somente
localização negócio quarto apenas achar
localização ótima localização maior qualidade
localização ótima piscina pleno avenida
localização perfeito travesseiro mínimo estranho
localização perto avenida paulista shopping
localização perto shopping bom prato
localização pesar favor bem receptivo
localização privilegiado próximo restaurante bom
localização privilegiado situado avenida paulista
localização próximo estação metrô vida
localização próximo metrô brigadeiro quadra
localização próximo paulista luxo dispor
localização quadra avenida paulista cercado
localização quadra avenida paulista hcor
localização quadra avenida paulista jardins
localização quadra shopping pátio paulista
localização serviço bom localização negar
localização serviço bom oferecer serviço
localização serviço melhorar tomado modelo
localização serviços limpeza manutenção item
localização valer/valar hospedagem bom café

Segundo a Tabela 5.12, o número máximo de palavras das proto-frases encontradas nas sentenças extraídas foi 4. Uma frase foi mapeada 3 vezes, mostrando-se mais representativa enquanto que outras três foram mapeadas uma vez cada uma. O conteúdo de todas as sentenças referem-se à localização do Hotel e detalha não apenas onde este se situa como também os lugares

Tabela 5.12: Frases extraídas após mapeamento da palavra "localização" indicando o número de palavras contidas na proto-frase e quantas vezes sua extração se repetiu.

N. de Palavras da proto-frase	**Repetições**	**Frase**
4	3	Hotel em excelente localização, a uma quadra da avenida paulista e próximo à estação de metrô brigadeiro.
4	1	Apesar de a localização pesar a favor por estar bem próximo da avenida paulista e as suas diversas estações de metrô, não merece nota mais alta , pois a cidade oferece opções melhores com maior qualidade na hospedagem.
4	1	Possuindo localização privilegiada, situado em avenida paralela à avenida paulista, próximo a vários restaurantes, posto de gasolina, americanas express, etc.
4	1	Excelente localização.

próximos. Em relação a opinião representativa deste atributo, destaca-se uma avaliação positiva como "excelente localização" e "localização privilegiada". Comparando estes resultados com a pesquisa da Figura 5.6, nota-se que a localização também possui uma avaliação positiva (quatro e meio de cinco pontos) corroborando com os achados do ECC.

Café

O terceiro atributo de maior grau é a palavra "café", a qual será o novo ponto de partida para executar o algoritmo de caminhos e o mapeamento. A Tabela 5.3.3 apresenta os parâmetros utilizados para o teste, a Tabela 5.14 as 45 proto-frases geradas e a Tabela 5.15 o resultado final do mapeamento, apresentando as frases extraídas, a quantidade de palavras da proto-frase que contém e quantas vezes foi mapeada pelas proto-frases.

Palavra	café
comprimento da proto-frase	5
número de proto-frases	45

Tabela 5.13: Parâmetros para execução do algoritmo de caminhos partindo da palavra "cafe".

Tabela 5.14: Proto-frases formadas pela execução do algoritmo de caminhos partindo da palavra “café”.

Proto-frases
café manhã jantar restaurante bom
café manhã bem localizar limpo
café manhã bom hotel três
café manhã quarto confortável ducha
café manhã variado próximo metrô
café restaurante passos paulista restaurante
café manhã variado inclusive omelete
café manhã maravilhoso variedade qualidade
café manhã ótimo café manhã
café manhã bem servido bom
café manhã bom hotel cobrar
café manhã excelente churrascaria próximo
café manhã razoável atendimento colaborador
café manhã correto roupa cama
café manha pecar creio diversidade
café manhã possuir ótimo hotel
café leite iogurte cereal fruta
café manhã correto roupa precisar
café manhã excelente café manhã
café manhã bom local café
café manhã quarto feitar/fazer quinze
café manha cinco minuto opinião
café manhã possuir piscina pleno
café manhã manhã bom tv
café manhã almoço jantar maravilhoso
café manhã bem devorado cama
café manhã possuir opção refeição
café manha farto variado inclusive
café manhã bom custo-benefício bem
café manhã quarto limpo café
café manhã razoável atendimento nota
café manhã manhã quatro bem
café manhã excelente trabalho passar
café manhã bem devorado cama
café restaurante bom termo cama
café manhã bom coração paulo
café manhã bastante diversidade qualidade
café manha bom local hotel
café manhã possuir opção restaurante
café manhã excelente localização excelente
café manha farto variado funcionar
café manhã correto roupa precisar
café manha bem próximo shopping

café manhã farto café manhã
café manhã impecável atendimento paulista

Tabela 5.15: Frases extraídas após mapeamento da palavra "café" indicando o número de palavras contidas na proto-frase e quantas vezes sua extração se repetiu.

N. de Palavras da proto-frase	**Repetições**	**Frase**
5	1	O café da manhã é variado, inclusive com omeletes feitos na hora.
5	1	O café da manhã é maravilhoso, com uma variedade de pães e bolos, omeletes feitos na hora, sucos da fruta, café, leite, iogurte, cereais, além de frutas.
5	1	Gostei muito do café da manhã, bastante diversidade e qualidade excelente.
5	1	Café da manha pecou um pouco , creio que uma diversidade maior de pães e guarnições seria ótimo.

Analisando os resultados apresentados na Tabela 5.15, foram extraídas 4 sentenças uma única vez contendo todas as palavras de suas respectivas proto-frases. Todas referem-se ao café-da-manhã e trazem ricos detalhes do que é servido. Sobre a avaliação deste atributo, duas frases classificam o café-da-manhã positivamente com adjetivos como "maravilhoso" e "qualidade excelente". Uma frase destaca a variedade e a última sentença aponta a necessidade de maior diversidade. No geral, a avaliação deste atributo é positiva. Porém, não é possível comparar estes resultados com a pesquisa quantitativa uma vez que este item não aparece na avaliação.

Restaurante

O quarto atributo de maior grau é a palavra "restaurante", o próximo ponto de partida para executar o algoritmo de caminhos e o mapeamento. A Tabela 5.16 apresenta os parâmetros utilizados para o teste, a Tabela 5.17 as 35 proto-frases geradas e a Tabela 5.18 o resultado final do mapeamento, apresentando as frases extraídas, a quantidade de palavras da proto-frase que contém e quantas vezes foi mapeada pelas proto-frases.

A Tabela 5.17 apresenta as 35 proto-frases geradas e a Tabela 5.18 o resultado final do mapeamento, apresentando as frases extraídas, a quantidade de palavras da proto-frase que contém e quantas vezes foi mapeada pelas proto-frases.

Tabela 5.16: Parâmetros para execução do algoritmo de caminhos partindo da palavra "restaurante".

Palavra	Restaurante
comprimento da proto-frase	5
número de proto-frases	35

Tabela 5.17: Proto-frases formadas pela execução do algoritmo de caminhos partindo da palavra "restaurante".

Proto-frases
restaurante all seasons desejar limpeza
restaurante bar bom localização perto
restaurante bar chefia chefe suíço
restaurante bar chefia chefe suíço
restaurante bares
restaurante bom bem legal cidade
restaurante bom café manhã diferencial
restaurante bom dia quatorze abril
restaurante bom ducha água vazar
restaurante bom nível conforto bom
restaurante bom ótimo hotel golden
restaurante bom qualidade cama extra
restaurante bom vizinhança perto avenida
restaurante casa bom nível paulo
restaurante cinemas metrô metrô parte
restaurante comer restaurante maravilhoso bom
restaurante comer restaurante regular máximo
restaurante continuar apenas começar perceber
restaurante especialmente ir paulo desprezar
restaurante especialmente viagem familiar acredito
restaurante hotel vontade passar quinze
restaurante magnifico assim melhoria item
restaurante manutenção qualidade assim golden
restaurante maravilhoso bom termo solicitado
restaurante notar vencido serviço recepção
restaurante ótimos restaurante caro então
restaurante possuir enorme tampo plástico
restaurante possuir piscina pleno região
restaurante posto gasolina americano express
restaurante posto gasolina americano express
restaurante prato quente barulho insuportável
restaurante remodelar então dias quente
restaurante requintado preço bom coração
restaurante shoppings loja centro negócio
restaurante variadas cozinha equipada

Segundo dados da Tabela 5.18, as duas sentenças extraídas apresentaram todas as palavras da proto-frase e foram mapeadas duas vezes. O conteúdo da primeira frase refere-se à qualidade do restaurante do hotel e detalha com o nome de quem o chefia. Já a segunda relaciona o termo

Tabela 5.18: Frases extraídas após mapeamento da palavra "restaurantes" indicando o número de palavras contidas na proto-frase e quantas vezes sua extração se repetiu.

N. de Palavras da proto-frase	**Repetições**	**Frase**
5	2	Possui ainda um ótimo restaurante e bar, sob a chefia do chefe suíço C. B.
5	2	Possuindo localização privilegiada, situado em avenida paralela à avenida paulista , próximo a vários restaurantes, posto de gasolina, americanas express, etc.

o item restaurante não foi avaliado.

Atendimento

O último atributo a ser avaliado, quinto maior grau dentre aqueles presentes na Tabela 5.4 é a palavra atendimento, ponto de partida para executar o algoritmo de caminhos e o mapeamento. A Tabela 5.19 apresenta os parâmetros utilizados para o teste, a Tabela 5.20 as 30 proto-frases geradas e a Tabela 5.21 o resultado final do mapeamento, apresentando as frases extraídas, a quantidade de palavras da proto-frase que contém e quantas vezes foi mapeada pelas proto-frases.

Tabela 5.19: Parâmetros para execução do algoritmo de caminhos partindo da palavra "atendimento".

Palavra	Atendimento
comprimento da proto-frase	5
número de proto-frases	30

Tabela 5.20: Proto-frases formadas pela execução do algoritmo de caminhos partindo da palavra "atendimento".

Proto-frases
atendimento atenção negativamente box apenas
atendimento bastante conforto ótimo café
atendimento bastante heterogêneo executivo turista
atendimento bom café-da-manhã refeição cansado
atendimento bom exceto cortina plástico
atendimento bom funcionalidade
atendimento cardápio almoço achar banheira
atendimento cliente estrangeiro impossível
atendimento colaborador higiene bom antar

atendimento colaborador higiene bom cozinha
atendimento excelente custo-benefício região paulista
atendimento excelente hotel bem limpo
atendimento excelente localização excelente serviço
atendimento excelente localização quarto recomer/recomendar
atendimento excelente opção ir balcão
atendimento excelente principal ponto gostar
atendimento funcionário atencioso solícito pronto
atendimento impecável ar condicionado poça
atendimento impecável equipe exceção amigável
atendimento necessita melhoria serviço bom
atendimento necessita melhoria serviço hotel
atendimento paulista proximidade paulista infraestrutura
atendimento paulista quarto barulhento dar
atendimento pontual sonorização climatização excelentes
atendimento pontual sonorização climatização excelentes
atendimento rápido faltar tomar banho
atendimento recepção errar escolha neste
atendimento restaurante manutenção item banheiro
atendimento serviço quarto bonito funcional
atendimento valer pena horário saída

Tabela 5.21: Frases extraídas após mapeamento da palavra "atendimento" indicando o número de palavras contidas na proto-frase e quantas vezes sua extração se repetiu.

N. de Palavras da proto-frase	**Repetições**	**Frase**
4	1	Estive hospedada em um final de semana, hotel maravilhoso, desde o atendimento na recepção no *check in* e *out*.
4	1	Você ter a sensação de bem estar como : uma acolhida ao se apresentar no *check in* , a liberação das suas acomodações com uma brevidade, o atendimento dos colaboradores, higiene, um bom café da manha.
4	1	Excelente localização e bom atendimento, mas necessita de melhorias em alguns itens.

Como é possível observar através da Tabela 5.21, as três sentenças extraídas possuem 4 palavras de suas respectivas proto-frases e foram extraídas uma única vez. O conteúdo refere-se de forma genérica ao atendimento, relatando alguns serviços do hotel como o *check-in* e *check-out*.

A única referência direta ao atendimento o avalia como "bom". As demais avaliações incluem o atendimento como um item de uma avaliação geral do hotel. Comparando estes resultados com a pesquisa quantitativa da Figura 5.6, o atendimento é avaliado com 4 pontos de 5, tendendo a uma avaliação mais próxima do bom do que do excelente, mesma percepção da terceira sentença extraída.

De forma geral, as sentenças extraídas possuem um conteúdo semântico muito próximo ao que era esperado segundo os dados da pesquisa presente na página do hotel, o que valida os resultados obtidos pelo ECC nos testes com este banco de dados em específico. A extração de frases, segundo a metodologia elaborada, mostrou-se um eficiente meio de se conhecer detalhadamente a percepção de um coletivo a respeito daquilo que é vivenciado por seus participantes. Por exemplo, foi possível conhecer não apenas que o hotel era bem localizado, mas também seu local exato e lugares próximos; ou então a qualidade do café-da-manhã foi descrita a partir daquilo o que é servido; ou ainda foi possível conhecer detalhes do quarto como seu espaço e até a decoração. Esta característica implica em vantagens quando o objetivo é desenvolver uma ferramenta para democracia digital uma vez que permite ir além das impressões gerais conhecendo com detalhes aquilo que uma população está querendo dizer. No próximo capítulo serão apresentadas as conclusões, focando os desafios no desenvolvimento do ECC como um instrumento de democracia participativa e suas perspectivas futuras.

Capítulo 6

Conclusões e Trabalhos Futuros

Desenvolver um projeto em democracia digital vai além de desafios puramente tecnológicos. Ferramentas para promoção da democracia só se tornam de fato efetivas se estiverem ao alcance de todos os cidadãos e despertarem o interesse da esfera política por maior participação e representatividade social. Durante a execução deste projeto de mestrado, um importante passo foi dado para o incentivo no desenvolvimento destas tecnologias no Brasil: Em 23 de maio de 2014, o governo brasileiro instituiu o decreto nº 8234 (BRASIL, 2014) que criou a Política Nacional de Participação Social (PNPS) cujo objetivo é fortalecer e articular os mecanismos e instâncias democráticas de diálogo e a atuação conjunta entre administração pública federal e a sociedade civil. Dentre suas diretrizes, aponta "o reconhecimento da participação social como direito do cidadão e expressão da sua autonomia" e a "valorização da educação para a cidadania ativa", abrindo portas para o resgate de uma cidadania participativa e consciente. Além disso, define importantes elementos de democracia participativa e uso da tecnologia. Destacam-se dois incisos do Artigo 2, os quais definem a consulta pública e o ambiente virtual, condições básicas para democracia digital:

IX - consulta pública - mecanismo participativo, a se realizar em prazo definido, de caráter consultivo, aberto a qualquer interessado, que visa a receber contribuições por escrito da sociedade civil sobre determinado assunto, na forma definida no seu ato de convocação; e

X - ambiente virtual de participação social - mecanismo de interação social que utiliza tecnologias de informação e de comunicação, em especial a internet, para promover o diálogo entre administração pública federal e sociedade civil.

Enquanto que o Artigo 3 aborda diretamente a questão do uso da tecnologia na participação social, apresentando como diretriz geral:

VI - incentivar o uso e o desenvolvimento de metodologias que incorporem múltiplas formas de expressão e linguagens de participação social, por meio da internet, com a adoção de tecnologias livres de comunicação e informação, especialmente, softwares e aplicações, tais como

códigos fonte livres e auditáveis, ou os disponíveis no Portal do Software Público Brasileiro;

A Política Nacional de Participação Social é uma importante conquista para o desenvolvimento de tecnologias que visam o aperfeiçoamento da democracia no Brasil, incentivando o surgimento de projetos tal como o presente mestrado. Porém, os desafios para o desenvolvimento destas tecnologias são inúmeros e o Extrator de Conhecimento Coletivo focou-se em abordar a questão do processamento da informação em massa como forma de dar voz ao coletivo.

A partir dos resultados obtidos pelo ECC, é possível concluir que o uso da teoria de redes complexas mostrou-se uma possível abordagem para o tratamento da informação coletiva. O banco de dados pôde ser mapeado como um sistema complexo com características típicas de redes mundo pequeno e comportamento que segue a lei da potência, indicando a presença de núcleos de informação e fácil navegabilidade. O Algoritmo de Caminhos, elaborado para navegar sobre a rede e buscar a informação coletiva, também mostrou-se uma eficiente heurística para extração dos parágrafos mais significativos. Porém, durante o desenvolvimento do ECC e a execução dos testes, algumas condições em relação ao pré-processamento e o mapeamento do banco de dados foram observadas e podem indicar caminhos para o aprimoramento dos resultados e maior eficiência da ferramenta.

Na etapa de pré-processamento, o qual envolve a lematização e retirada das *stop-words*, foi verificado que quanto menos palavras no texto final melhor é a qualidade da rede no que diz respeito às características mundo pequeno e o comportamento livre de escala. Uma possível abordagem para minimizar a quantidade de palavras seria utilizar um dicionário de sinônimos para padronizar alguns termos sem modificar o significado original das sentenças. Desta forma, sentenças de conteúdo semântico semelhantes tenderão a se aglomerar nos mesmos ramos da rede, facilitando a emergência da informação coletiva. Na etapa de mapeamento da proto-frase no texto lematizado, as sentenças são selecionadas segundo o número de palavras iguais, condição a qual resulta em uma significativa quantidade de frases diferentes extraídas, conforme observado nos resultados dos testes. Uma possibilidade para melhorar a eficiência deste mapeamento seria considerar não apenas a quantidade de palavras repetidas, mas também a ordem em que elas aparecem no texto lematizado. Para aprimoramentos futuros, sugere-se também testar outras métricas de centralidade ou até utilizar algoritmos de aglomeração na seleção dos núcleos da rede.

Outro aspecto a ser observado é em relação à validação dos resultados obtidos. O único instrumento utilizado para verificação da representatividade dos resultados foi a pesquisa quantitativa presente na página do hotel. Porém, para trabalhos futuros, seria interessante o uso de métodos mais robustos de validação como uso de juízes de prova, onde avaliadores selecionariam as frases que julgam mais representativas após lerem todos os depoimentos e os resultados seriam comparados com as escolhas do ECC. Além disso, outra limitação foi a própria constituição do banco de dados, o qual não diz respeito diretamente a aplicabilidade do instrumento. Para testes posteriores, seria desejável escolher uma comunidade e aplicar uma pesquisa com o propósito de recolher os relatos dos participantes.

6.1 Perspectivas futuras

Além de aprimorar processos de mineração de dados, as perspectivas futuras para o Extrator de Conhecimento Coletivo envolvem o desenvolvimento de uma plataforma virtual de inteligência coletiva, inspirada na ideia da Ágora Virtual de Lé6vy (1999), onde os participantes de uma comunidade podem livremente se expressar, conhecer e resolver problemas de sua realidade. A proposta é que o ECC seja o núcleo de processamento de dados desta plataforma onde tornará expresso, em tempo real, o conteúdo que melhor representa aquela comunidade. Para que isto seja possível, o ECC deve se tornar uma ferramenta virtual e deve ser desenvolvida uma interface humano-computador amigável e acessível. Estes desafios farão parte do projeto de doutorado oriundo deste trabalho.

Por fim, conclui-se que desenvolver ferramentas para democracia digital não é apenas um desafio de ordem tecnológica, mas também social e político que resgata a necessidade de se pensar a engenharia de forma multidisciplinar e refletir os impactos da produção cientifico-tecnológica no bem estar-social. Para que projetos de democracia participativa sejam bem sucedidos, os compromissos vão além do meio acadêmico, pois é preciso mais do que metodologias, recursos tecnológicos ou ferramentas participativas para que o exercício da cidadania possa interferir positivamente na vida de uma comunidade: questões como a inclusão digital, igualdade no acesso à informação, liberdade de expressão, a educação política, bem como a conscientização dos cidadãos sobre as questões políticas em jogo, são elementos fundamentais e conquistas não menos importantes para a qualidade do processo democrático.

Bibliografia

ABELLO, J.; BUCHSBAUM, A. L.; WESTBROOK, J. R. A functional approach to external graph algorithms. In: *6th ESA*. Venicy: Springer-Verlag, 1998. p. 332–343.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics*, APS, v. 74, n. 1, p. 47–97, 2002.

AMARAL, L. A.; OTTINO, J. M. Complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, Springer, v. 38, n. 2, p. 147–162, 2004.

ANDRITSOS, P. Data clustering techniques. *Toronto, University of Toronto, Dep. of Computer Science*, v. 1, n. 1, p. 34, 2002.

ANTIQUEIRA, L.; JR, O. N. O.; COSTA, L. d. F.; NUNES, M. d. G. V. A complex network approach to text summarization. *Information Sciences*, Elsevier, v. 179, n. 5, p. 584–599, 2009.

ANTIQUEIRA, L.; NUNES, M. d. G. V.; JR, O. O.; COSTA, L. d. F. Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 373, p. 811–820, 2007.

ANTIQUEIRA, L.; PARDO, T. A. S.; NUNES, M. d. G. V.; JR, O. N. O.; COSTA, L. d. F. Some issues on complex networks for author characterization. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, v. 11, n. 36, p. 51–58, 2007.

ARISTOTELES. *Aristotle's Politics*. Des Moines: The Peripatetic Press, 1986 (c.320 BC). C.320 BC.

BALINSKY, A.; BALINSKY, H.; SIMSKE, S. On the helmholtz principle for data mining. *Hewlett-Packard Development Company, LP*, 2011.

BALINSKY, H.; BALINSKY, A.; SIMSKE, S. J. Automatic text summarization and small-world networks. In: *Proceedings of the 11th ACM symposium on Document engineering*. New York, NY, USA: ACM, 2011. (DocEng '11), p. 175–184.

BARABÁSI, A.-L.; ALBERT, R.; JEONG, H. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 272, n. 1, p. 173–187, 1999.

BARBER, B. R. *Strong Democracy: Participatory politics for a new age*. Oakland: University of California Press, 2003.

BATAGELJ, V.; MRVAR, A. Pajek-program for large network analysis. *Connections*, v. 21, n. 2, p. 47–57, 1998.

BEITZEL, S. M.; JENSEN, E. C.; LEWIS, D. D.; CHOWDHURY, A.; FRIEDER, O. Automatic classification of web queries using very large unlabeled query logs. *ACM Transactions on Information Systems*, ACM, v. 25, n. 2, p. 9, 2007.

BEN-NAIM, E.; FRAUENFELDER, H.; TOROCZKAI, Z. *Complex networks*. Berlim: Springer, 2004.

BENNETT, W. L.; ENTMAN, R. M. *Mediated politics: Communication in the future of democracy*. Cambridge: Cambridge University Press, 2001.

BISHOP, C. M. *Neural networks for pattern recognition*. Oxford: Oxford university press, 1995.

BOLLOBÁS, B. *Random graphs*. Berlim: Springer, 1998.

BOLLOBÁS, B.; RIORDAN, O. The diameter of a scale-free random graph. *Combinatorica*, Springer, v. 24, n. 1, p. 5–34, 2004.

BRABHAM, D. C. Crowdsourcing the public participation process for planning projects. *Planning Theory*, Sage Publications, v. 8, n. 3, p. 242–262, 2009.

BRASIL. *Decreto nº 8.243, de 23 de maio de 2014*. 2014. Institui a Política Nacional de Participação Social - PNPS e o Sistema Nacional de Participação Social - SNPS, e dá outras providências. Diário Oficial [da República Federativa do Brasil], Brasília, seção 1, página 6, 26 maio de 2014.

BRODER, A.; KUMAR, R.; MAGHOUL, F.; RAGHAVAN, P.; RAJAGOPALAN, S.; STATA, R.; TOMKINS, A.; WIENER, J. Graph structure in the web. *Computer networks*, Elsevier, v. 33, n. 1, p. 309–320, 2000.

BUCY, E. P.; GREGSON, K. S. Media participation a legitimizing mechanism of mass democracy. *New media & society*, SAGE Publications, v. 3, n. 3, p. 357–380, 2001.

BULLMORE, E.; SPORNS, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, Nature Publishing Group, v. 10, n. 3, p. 186–198, 2009.

CANCHO, R. F. i; SOLÉ, R. V. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, The Royal Society, v. 268, n. 1482, p. 2261–2265, 2001.

CANFORA, L. *Democracy in Europe: A History of an Ideology*. New York: John Wiley & Sons, 2008.

CARVALHO, M. G. d. Tecnologia, desenvolvimento social e educação tecnológica. *Revista Educação & Tecnologia*, n. 1, p. 14, 1997.

CASTANHO, M. A. F. d. S. Internet como instrumento de revitalização da representação política. *Revista Democracia Digital e Governo Eletrônico*, n. 6, p. 20–45, 2012.

CHANDLER, D. *Introduction to modern statistical mechanics*. Oxford: Oxford University Press, 1987.

CLIFT, S. *E-democracy, e-governance and public network*. 2003. Disponível em: <http://www.opensourcejahrbuch.de/Archiv/2005/2005/abstracts/2004/pdfs/IV-5-Clift.pdf>.

COLEMAN, S.; BLUMLER, J. G. *The Internet and democratic citizenship: Theory, practice and policy*. Cambridge: Cambridge University Press, 2009.

CUNNINGHAM, F. *Theories of democracy: a critical introduction*. Cambridge: Cambridge University Press, 2002.

DAHLBERG, L. Re-constructing digital democracy: An outline of four 'positions'. *New Media & Society*, Sage Publications, v. 13, n. 6, p. 855–872, 2011.

DIESTEL, R. *Graph Theory*. New York: Springer-Verlag Berlin and Heidelberg GmbH & Company KG, 2000.

DIJK, J. V. Models of democracy and concepts of communication. *Digital democracy: Issues of theory and practice*, Sage London, p. 30–53, 2000.

DINIZ, V. A história do uso da tecnologia da informação na gestão pública brasileira através do CONIP–congresso de informática pública. In: *X CONGRESO INTERNACIONAL DEL CLAD SOBRE LA REFORMA DEL ESTADO Y DE LA ADMINISTRACIÓN PÚBLICA*. Santiago, Chile: CLAD, 2005.

DUBES, R. C.; JAIN, A. K. Clustering methodologies in exploratory data analysis. *Advances in computers*, v. 19, n. 11, p. 113–228, 1980.

DUMAIS, S. T. Latent semantic analysis. *Annual review of information science and technology*, Wiley Online Library, v. 38, n. 1, p. 188–230, 2004.

ERDŐS, P.; RÉNYI, A. On random graphs. *Publicationes Mathematicae Debrecen*, v. 6, p. 290–297, 1959.

ERKAN, G.; RADEV, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, v. 22, n. 1, p. 457–479, 2004.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. *Computer Communication Review*, ACM, v. 29, n. 4, p. 251–262, 1999.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. *Advances in knowledge discovery and data mining*. Cambridge: the MIT Press, 1996.

FEENBERG, A. *Critical theory of technology*. New York: Oxford University Press, 1991.

GOMES, W. A democracia digital e o problema da participação civil na decisão política. *Fronteiras-estudos midiáticos*, v. 7, n. 3, p. 214–222, 2005.

GOMES, W. Internet e participação política em sociedades democráticas. *Revista FAMECOS: mídia, cultura e tecnologia*, v. 1, n. 27, p. 58–78, 2006.

GRONLUND, A.; HORAN, T. A. Introducing e-gov: History, definitions, and issues. *Communications of the Association for Information Systems*, v. 15, p. 713–729, 2005.

GUPTA, V.; LEHAL, G. S. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, v. 2, n. 3, p. 258–268, 2010.

HALL, S. *A identidade cultural na pós-modernidade*. Rio de Janeiro: DPA, 2006.

HALTEREN, H. V. *Writing style recognition and sentence extraction*. 2002. Disponível em: <http://www-nlpir.nist.gov/projects/duc/pubs/2002papers/nijmegen_vanhalteren.pdf>.

HELD, D. *Models of democracy*. Cambridge: Polity, 2006.

HIRAO, T.; SASAKI, Y.; ISOZAKI, H.; MAEDA, E. Ntts text summarization system for duc-2002. In: CITESEER. *Proceedings of the Document Understanding Conference 2002*. Philadelphia, 2002. p. 104–107.

HUBERMAN, B. A.; ADAMIC, L. A. Internet: growth dynamics of the world-wide web. *Nature*, Nature Publishing Group, v. 401, n. 6749, p. 131–131, 1999.

JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABÁSI, A.-L. The large-scale organization of metabolic networks. *Nature*, Nature Publishing Group, v. 407, n. 6804, p. 651–654, 2000.

JIANG, B.; CLARAMUNT, C. Topological analysis of urban street networks. *Environment and Planning B*, PION LTD, v. 31, n. 1, p. 151–162, 2004.

JONES, K. S. Automatic summarising: The state of the art. *Information Processing & Management*, Elsevier, v. 43, n. 6, p. 1449–1481, 2007.

KAIKHAH, K. Automatic text summarization with neural networks. In: *Second International IEEE Conference on Intelligent Systems Proceedings*. Varna, Bulgária: IEEE, 2004. p. 40–45.

KIREYEV, K. Using latent semantic analysis for extractive summarization. In: *Proceedings of text analysis conference*. Gaithersburg, Maryland, USA: NIST, 2008.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, ACM, v. 46, n. 5, p. 604–632, 1999.

KOCHEN, M. *The small world*. New Jersey: Ablex Norwood, 1989.

KONTOSTATHIS, A.; GALITSKY, L. M.; POTTENGER, W. M.; ROY, S.; PHELPS, D. J. *A survey of emerging trend detection in textual data mining*. 2004. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1B868BB1E216D2530A7282A07A0678DB?doi=10.1.1.19.4190&rep=rep1&type=pdf>.

LAVALLE, A. G.; HOUTZAGER, P. P.; CASTELLO, G. Democracia, pluralização da representação e sociedade civil. *Lua Nova*, SciELO Brasil, v. 67, n. 67, p. 49–103, 2006.

LÉVY, P. *Collective intelligence: Mankind's emerging world in cyberspace*. Jackson: Perseus Publishing, 1999.

LÉVY, P. *Cyberdémocratie*. Paris: Odile Jacob, 2002.

LIDÉN, G. Supply of and demand for e-democracy: A study of the swedish case. *Information Polity*, IOS Press, v. 18, n. 3, p. 217–232, 2013.

LILJEROS, F.; EDLING, C. R.; AMARAL, L. A. N.; STANLEY, H. E.; ÅBERG, Y. The web of human sexual contacts. *Nature*, Nature Publishing Group, v. 411, n. 6840, p. 907–908, 2001.

MANI, I. *Automatic summarization*. Philadelphia: John Benjamins Publishing, 2001.

MARADEI, A. Folha de S. Paulo e a cobertura dos protestos do MPL. In: *XXXVI Congresso Brasileiro de Ciências da Comunicação*. Manaus: INTERCOM, 2013.

MASLOV, S.; SNEPPEN, K. Specificity and stability in topology of protein networks. *Science*, American Association for the Advancement of Science, v. 296, n. 5569, p. 910–913, 2002.

MATHIOUDAKIS, M.; KOUDAS, N. Twittermonitor: trend detection over the twitter stream. In: *2010 ACM SIGMOD International Conference on Management of data*. Indianapolis, Indiana, USA: ACM, 2010.

MENEZES, M. L. D. Democracia de assembleia e democracia de parlamento: uma breve história das instituições democráticas. *Sociologias*, SciELO Brasil, v. 12, n. 23, p. 20–45, 2010.

MEZARROBA, M. P.; JUNIOR, E. S.; ALVES, J. B. D. M.; ROVER, A. J. O portal e-democracia da câmara dos deputados como sistema sócio-tecnológico. *Revista Democracia Digital e Governo Eletrônico*, n. 9, p. 24–43, 2013.

MIGUEL, L. F. *Representação política em 3-D: elementos para uma teoria ampliada da representação política*. [S.l.]: SciELO Brasil, 2006. 123–140 p.

MIHALCEA, R.; TARAU, P. *TextRank: Bringing order into texts*. 2004. Disponível em: <http://digital.library.unt.edu/ark:/67531/metadc30962/>.

MILGRAM, S. The small world problem. *Psychology today*, New York, v. 2, n. 1, p. 60–67, 1967.

MILL, J. S. Considerations on representative government. *inCollected Works of John Stuart Mill, ed. RobsonJ. M.(Toronto: University of Toronto Press, 1977)*, p. 572–73, 1861.

MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, ACM, v. 38, n. 11, p. 39–41, 1995.

MITCHELL, M. Complex systems: Network thinking. *Artificial Intelligence*, Elsevier, v. 170, n. 18, p. 1194–1212, 2006.

MONTOYA, J. M.; SOLÉ, R. V. Small world patterns in food webs. *Journal of theoretical biology*, Elsevier, v. 214, n. 3, p. 405–412, 2002.

MOTTER, A. E.; MOURA, A. P. de; LAI, Y.-C.; DASGUPTA, P. Topology of the conceptual network of language. *Physical Review E*, APS, v. 65, n. 6, p. 065102, 2002.

NARAYANAMURTI, V.; ODUMOSU, T.; VINSEL, L. *The Discovery-Invention Cycle: Bridging the Basic/Applied Dichotomy*. 2013. Disponível em: <http://belfercenter.ksg.harvard.edu/files/narayanamurti_odumosu_vinsel_2013_dp.pdf>.

NCHISE, A. C. The trend of e-democracy research: summary evidence and implications. In: *13th Annual International Conference on Digital Government Research*. College Park, MD, USA: ACM, 2012.

NEWMAN, M. *Networks: an introduction*. New York: Oxford University Press, 2010.

NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. *The structure and dynamics of networks*. Princeton: Princeton University Press, 2006.

PACHECO, D. C.; REGINALDO, T.; FRANZONI, A. M. B.; BALDESSAR, M. J. Governança eletrônica e inclusão digital na prefeitura municipal de criciúma (sc). *Revista Democracia Digital e Governo Eletrônico*, n. 9, p. 101–123, 2013.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. *The PageRank citation ranking: Bringing order to the web*. 1999. Disponível em: <http://ilpubs.stanford.edu:8090/422/>.

PASTOR-SATORRAS, R.; VESPIGNANI, A. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge: Cambridge University Press, 2007.

PERISSINOTTO, R. M.; FUKS, M. *Democracia: teoria e prática*. Rio de Janeiro: Relume Dumará, 2002.

PHARR, S. J.; PUTNAM, R. D. *Disaffected democracies: what's troubling the trilateral countries?* Princeton: Princeton University Press, 2000.

PITKIN, H. *The Conception of Representation*. London: University of California Press, 1971.

RAMOS, J. *Using TF-IDF to determine word relevance in document queries*. 2003. Disponível em: <https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf>.

REDNER, S. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, Springer, v. 4, n. 2, p. 131–134, 1998.

RIBEIRO, P.; SOPHIA, D. C.; GRIGÓRIO, D. d. A. Gestão governamental e sociedade: informação, tecnologia e produção científica. *Ciênc. saúde coletiva*, SciELO Public Health, v. 12, n. 3, p. 623–31, 2007.

ROMAN, A. V.; MILLER, H. T. New questions for e-government: Efficiency but not (yet?) democracy. *International Journal of Electronic Government Research (IJEGR)*, IGI Global, v. 9, n. 1, p. 65–81, 2013.

SALGANIK, M. J.; LEVY, K. E. *Wiki surveys: Open and quantifiable social data collection*. 2012. Disponível em: <http://arxiv.org/pdf/1202.0500v1.pdf>.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, ACM, v. 34, n. 1, p. 1–47, 2002.

SIGMAN, K. Appendix: A primer on heavy-tailed distributions. *Queueing Systems*, Springer, v. 33, n. 1, p. 261–275, 1999.

SIGMAN, M.; CECCHI, G. A. Global organization of the wordnet lexicon. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 99, n. 3, p. 1742–1747, 2002.

SILVA, A. K. Araújo da; CORREIA, A. E. G. C.; LIMA, I. França de. O conhecimento e as tecnologias na sociedade da informação. *Revista Interamericana de Bibliotecología*, v. 33, n. 1, p. 213–239, 2010.

SILVA, S. P. d. Graus de participação democrática no uso da internet pelos governos das capitais brasileiras. *Opinião Pública*, SciELO Brasil, v. 11, p. 450–468, 2005.

SPIRAKIS, G.; SPIRAKI, C.; NIKOLOPOULOS, K. The impact of electronic government on democracy: e-democracy through e-participation. *Electronic Government, an International Journal*, Inderscience, v. 7, n. 1, p. 75–88, 2010.

STEMMER. Disponível em:. *http://www.nilc.icmc.usp.br/nilc/tools/stemmer.html. Acessado em Junho de 2013*, 2013.

SUANMALI, L.; BINWAHLAN, M. S.; SALIM, N. Sentence features fusion for text summarization using fuzzy logic. In: *Ninth International Conference on*. Shenyang, LiaoNing, China: IEEE, 2009.

THAKKAR, K. S.; DHARASKAR, R. V.; CHANDAK, M. Graph-based algorithms for text summarization. In: *3rd International Conference on Emerging Trends in Engineering and Technology (ICETET)*. Goa, India: IEEE, 2010.

TIMONEN, A. Digital democracy in the EU. *European View*, Springer, v. 12, n. 1, p. 103–112, 2013.

VEDEL, T. The idea of electronic democracy: Origins, visions and questions. *Parliamentary Affairs*, Hansard Soc, v. 59, n. 2, p. 226–235, 2006.

WANG, X. F.; CHEN, G. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE*, IEEE, v. 3, n. 1, p. 6–20, 2003.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

WEISS, S. M. *Predictive data mining: a practical guide*. Burlington: Morgan Kaufmann, 1998.

WEST, D. B. et al. *Introduction to graph theory*. Upper Saddle River: Prentice hall, 2001.

WILDAUER, E. W.; INABA, T. M. M.; SILVA, G. P. da. A distribuição da internet nos domicílios brasileiros e suas perspectivas futuras. *Revista Democracia Digital e Governo Eletrônico*, n. 9, p. 124–137, 2013.

WU, V. Gabinete digital: Metodologias inovadoras em consultas públicas online. In: *VI CONGRESSO DE GESTÃO PÚBLICA*. Brasilia: Conselho Nacional de Secretários de Estado de Administração, 2013.

YOUNG, I. M. Representação política, identidade e minorias. *Lua Nova*, SciELO Brasil, v. 67, p. 139–190, 2006.