XI Encontro Nacional de Pesquisa em Ciência da Informação
Inovação e inclusão social: questões contemporâneas da informação
Rio de Janeiro, 25 a 28 de outubro de 2010

GT 8 – Informação e Tecnologia
Modalidade de apresentação: Comunicação Oral

UMA ABORDAGEM BASEADA EM MÉTRICAS DE REDES COMPLEXAS PARA O ESTABELECIMENTO DO GRAU DE INFLUÊNCIA DE TERMOS EM DOCUMENTOS

Wladimir Cardoso Brandão
Universidade Federal de Minas Gerais
Fernando Silva Parreiras
University of Koblenz-Landau, Alemanha

Resumo: Nos últimos anos, a área de recuperação de informação tem recebido atenção especial da comunidade científica mundial. Pesquisas relacionadas à melhoria de métodos e algoritmos para recuperação de informação textual tem se ampliado, concentradas, em grande parte, no aprimoramento do modelo vetorial, em especial na busca por métodos e funções mais eficientes para cálculo de similaridade entre documentos e consultas. Paralelamente, a análise de redes complexas tem despertado o interesse da comunidade científica devido a sua capacidade de representação de problemas complexos de maneira objetiva, oferecendo um arcabouço teórico e prático para o estudo das propriedades e comportamentos dos elementos e relações que compõem os problemas. Recentemente, pesquisas considerando documentos como redes complexas de palavras vem sendo desenvolvidas. Entretanto, as possibilidades de utilização desta abordagem na resolução de problemas de recuperação e classificação de informação ainda foram pouco exploradas. O presente artigo apresenta uma abordagem baseada em métricas de redes complexas para obtenção de uma função de atribuição de pesos a termos em documentos. A presente abordagem apresentou precisão equivalente ao modelo vetorial quando aplicada para estimativa de similaridade entre documentos e consultas a partir de uma coleção de referência, o que evidencia a aplicabilidade de métricas de redes complexas de palavras em problemas de recuperação de informação.
Palavras-chave: Recuperação de informação, redes complexas, redes de palavras, métricas de rede.

1. INTRODUÇÃO

Apesar dos avanços no campo de recuperação de informação, em especial na melhoria dos métodos e algoritmos para cálculo de similaridade entre documentos e consultas e para ordenação de resultados, a análise de redes complexas como abordagem para melhorar o desempenho dos sistemas de recuperação e de classificação de informação foi pouco explorada.

Apenas recentemente, em grande medida a partir dos estudos de Kleinberg (1999, 2000a, 2000b), o tema redes complexas começou a despertar de maneira mais intensa o interesse da comunidade científica mundial. Embora em pequeno número, esforços procurando adotar o arcabouço teórico e prático provenientes da análise de redes complexas na tentativa de obtenção de maior desempenho em processos de sumarização, recuperação e classificação de informação textual têm sido observados nos últimos anos. Nesse sentido, pesquisadores assumem como hipótese que o conhecimento da estrutura, do comportamento e das propriedades de redes complexas de documentos, de termos, de co-autoria e de citação potencializa o desenvolvimento de técnicas e algoritmos mais eficientes a serem aplicados na resolução de problemas relacionados à organização e ao tratamento de informação.

O presente artigo apresenta uma abordagem baseada em métricas de redes complexas para obtenção de funções mais eficientes para atribuição de pesos a termos em documentos. Apresenta também os resultados experimentais da aplicação desta abordagem em um problema de cálculo de similaridade entre documentos e consultas. A abordagem apresentou desempenho equivalente ao esquema $\boldsymbol{TF-IDF}$[1] utilizado pelo modelo vetorial, considerando como métricas de avaliação a precisão[2] e a revocação[3], o que demonstra o potencial de utilização de métricas de redes complexas de palavras em problemas de recuperação de informação.

Na sessão 2 são apresentados trabalhos relacionados. A sessão 3 descreve alguns conceitos básicos da teoria de redes complexas necessários para a compreensão da abordagem. A sessão 4 apresenta medidas de influência associadas a cálculos de similaridade. A sessão 5 descreve a abordagem proposta para obtenção de novas funções de atribuição de pesos. A

[1] Esquemas TF-IDF consideram as freqüências dos termos nos documentos (TF) e as freqüências invertidas dos termos na coleção (IDF) para calcular pesos de termos em documentos.
[2] Precisão é uma medida de exatidão usada para estimar o desempenho de um sistema de recuperação de informação.
[3] Revocação é uma medida de completude usada para estimar o desempenho de um sistema de recuperação de informação.

sessão 6 descreve o experimento efetuado, bem como apresenta os resultados experimentais obtidos. Finalmente, a sessão 7 conclui o artigo apontando direções para trabalhos futuros.

2. TRABALHOS RELACIONADOS

Burgess *et al.* (2006), ao procurarem estabelecer uma nova métrica indicativa da influência de um nó em um grafo, baseando-se apenas em aspectos topológicos, determinam como satisfatória a métrica denominada *eigenvector centrality* (BONACICH, 1972). A expectativa dos autores é de que tal métrica possa ser utilizada como uma alternativa mais eficiente aos métodos tradicionais de ordenamento de páginas na *Web* utilizadas pelos mecanismos de busca.

Korfiatis *et al.* (2007) descrevem modelos que provêem a integração de técnicas de ordenamento de documentos para recuperação de informação e técnicas de análise de rede social. A julgar pela tendência dos sistemas baseados em computador absorverem cada vez mais conceitos das áreas sociais, como reputação e credibilidade, tal integração será ampliada com o decorrer do tempo. Abordagens que consideram documentos como cadeias de conceitos são descritas por Srihari *et al.* (2005) como alternativas aos métodos tradicionais de recuperação de informação, permitindo a representação mais sofisticada de consultas e documentos e a construção de algoritmos igualmente sofisticados que implementem técnicas avançadas de mineração em grafos. Seguindo a mesma linha, Montes-Y-Gómez *et al.* (2000) apresentam um modelo para recuperação de informação e mineração de textos baseado em medidas de similaridade entre grafos representativos de sentenças e documentos.

Brandes *et al.* (2006), a fim de lidar com o problema da sumarização automática de textos em coleções heterogêneas, propõem um modelo para identificação de similaridades e dos graus de influência de documentos e termos em uma coleção. Tal modelo baseou-se numa estrutura de rede bipartida de termos e documentos e utilizou um processo analítico denominado *spectral analysis*. Como medida de influência, os autores testaram diferentes tipos de métricas, como *tf-idf* e *betweenness centrality*, sendo que a última demonstrou ser mais eficiente para capturar um número maior de informações estruturais dos documentos. Xie (2005) demonstra que o uso de métricas de redes complexas, utilizadas em conjunto ou separadamente das métricas de influência tradicionais do campo da recuperação de informação pode tornar o processo de sumarização mais eficiente.

Outra abordagen (ERKAN; RADEV, 2004a) leva em consideração métricas de prestígio, como *eigenvector centrality*, para filtrar os mais importantes sintagmas[4] em um documento com o intuito de melhorar a eficiência do processo de sumarização. Um método baseado em grafos para determinação de prestígio e influência em processamento de termos em linguagem natural é descrito detalhadamente por Erkan e Radey (2004b). Na comparação entre as métricas feitas pelos autores, as baseadas em centralidade de rede se mostraram melhores que as outras comumente utilizadas para resolução de problemas de sumarização.

Brandes e Cornelsen (2001) propõem um método de recuperação de informação baseado na visualização de documentos que suporta, simultaneamente, a exploração de sua estrutura interna de links e o seu ordenamento relativo na coleção. Tal método utiliza-se de métricas como *eigenvector centrality* e *outdegree centrality* para a obtenção da função de ordenamento. Apesar de demonstrar utilidade na exploração de estruturas inter-relacionadas de documentos, o método provou ser ineficiente para redes complexas.

Ainda considerando métodos que exploram a estrutura de links de documentos, Huang e Lai (2003) propõem a utilização de uma combinação de métricas de rede (*degree centrality*, *betweenness centrality* e *closeness centrality*) para imputação de importância e relevância em uma rede de documentos. Segundo os autores, o ordenamento de importância obtido a partir de seu método pode ser associado a perfis de usuários em um sistema de disseminação seletiva de informação, a fim de determinar com maior precisão o grau de relevância de um documento para o usuário.

Chitrapura e Kashyap (2004) propõem um método de ordenamento dinâmico de documentos mais eficaz que os métodos tradicionais que usam *PageRank*[5]. Tal método se baseia na associação de fluxos de valores, calculados a partir da métrica de *outdegree centrality*, aos nós da rede para determinação do grau de influência do nó em relação à rede como um todo. Tal método apresentou baixo impacto no tempo de execução da consulta apresentado melhores resultados que os métodos tradicionais.

Kurland e Lee (2005) apresentam um método de ordenamento da lista de documentos retornados por uma consulta através do estabelecimento de um nova ordenação a partir da

[4] Sintagmas são conjuntos de termos que carregam significado.
[5] *PageRank* é um método para atribuição de pesos a documentos que leva em conta as conexões entre os documentos da coleção.

exploração das relações assimétricas entre os elementos do conjunto retornado. Utilizando métricas de centralidade de rede como critério para o estabelecimento de novas funções de ordenamento, os autores demonstraram que o método melhorou a precisão na recuperação dos primeiros 10 documentos retornados em 10%.

Zhou *et al.* (2006) propõem um modelo para recuperação de informação de bases textuais baseado no uso das relações semânticas entre termos obtidas através da extração de sintagmas dos documentos e seu posterior casamento e substituição por termos presentes em redes semânticas. Tal modelo apresentou ganhos significativos de precisão (27%) na recuperação dos 100 primeiros documentos retornados para uma consulta.

3. REDES COMPLEXAS

O conceito de rede é aplicado em diversas áreas do conhecimento humano. Genericamente pode-se definir uma rede como um conjunto de elementos que mantêm conexões uns com os outros. Na literatura matemática, as redes são reconhecidas como grafos, seus elementos como vértices e suas conexões como arestas. Já nas ciências sociais, os elementos são denominados atores e as conexões são laços. Por outro lado, na literatura da ciência da computação, os elementos são reconhecidos como nós e as conexões como ligações.

No mundo real, sistemas podem ser representados e problemas podem ser tratados através da abordagem de rede. Um grupo de pessoas em uma organização trocando mensagens eletrônicas a fim de desempenhar suas funções pode ser interpretado como uma rede social, onde cada pessoa passa a ser um ator e as mensagens eletrônicas por eles trocadas passam a ser os laços da rede. Nesse sentido, o entendimento das redes, de sua estrutura, propriedades e comportamento, é fundamental para a compreensão das diversas classes de sistemas e problemas que podem ser por elas modelados e tratados.

Em redes simples, com dezenas ou centenas de nós e ligações, a própria visualização e interpretação do grafo a olho nu se constitui em importante ferramenta de análise. Entretanto, a modelagem da grande maioria dos sistemas e problemas reais envolve redes complexas, com milhares, milhões, ou mesmo bilhões de nós e ligações. Além disso, os nós em redes complexas podem assumir diferentes formas e apresentarem diferentes atributos, e as ligações podem ter significados diferentes podendo assumir valores e terem orientação. Para essa clas-

se de rede, a análise a olho nu se torna de pouca valia, uma vez que a quantidade de informação é tão grande que inviabiliza sua completa visualização, o que torna praticamente impossível seu processamento visual pelo cérebro humano.

3.1.Modelos e Redes Reais

Uma gama de problemas do mundo real podem ser modelados como redes complexas. A internet, a malha rodoviária e ferroviária e o sistema de distribuição de energia de um país podem ser interpretados como redes tecnológicas complexas. A *Web*, assim como as redes de citação entre artigos acadêmicos e os *thesaurus*[6] podem ser entendidos como complexas redes de conhecimento uma vez que sua estrutura reflete a estrutura de armazenamento de informação em seus elementos. Sistemas biológicos também podem ser representados por redes: um exemplo é a cadeia predatória entre animais onde os nós representam as diversas classes de animais presentes na fauna e as ligações representam uma relação de predação entre duas classes.

Ao longo dos anos modelos matemáticos foram desenvolvidos visando prover métodos e mecanismos para análise de redes complexas. Modelos de geração e crescimento de redes têm sido propostos e suas propriedades têm sido estudadas. Dentre as propriedades cabe destacar as relacionadas ao tamanho da rede, como o diâmetro (*diameter*), aos graus de centralidade dos nós, tais como o grau de centralidade (*degree centrality*), o grau de intermediação (*beetweenness centrality*) e o grau de proximidade (*closeness centrality*), ao grau de transitividade, tal como o coeficiente de agrupamento (*clustering coefficient*) e às suas respectivas distribuições estatísticas.

Modelos de geração criam redes que apresentam características particulares no que tange algumas de suas propriedades. Erdös e Rényi (1959, 1960) propuseram a geração de uma rede a partir de ligações estabelecidas de maneira aleatória entre seus nós, ou seja, todos os nós da rede têm a mesma probabilidade de estabelecerem relações uns com os outros. Tal modelo acaba por gerar uma rede denominada rede randômica ou aleatória, com a característica peculiar de apresentar diâmetro pequeno e baixo coeficiente de agrupamento.

---

[6] *Thesaurus* são vocabulários controlados freqüentemente utilizados na indexação e rotulação de documentos.

Watts e Strogatz (1998), baseando-se no famoso experimento de Stanley Milgram da década de 1960, denominado posteriormente por Guare (1990) de *"Six Degrees of Separation"*, propõem um modelo de geração de rede a partir da reescrita ou adição de um pequeno número de ligações de maneira aleatória em uma rede regular[7], o que acaba por gerar uma rede do tipo mundo-pequeno (*small-world*) com a característica peculiar de apresentar pequeno diâmetro e elevado coeficiente de agrupamento.

No modelo de crescimento proposto por Barabási e Albert (1999), nós entrantes na rede se associam as nós presentes por regras de preferência (*preferential attachment*), sendo que os nós com maior número de ligações têm maior probabilidade de receber novas ligações que os outros. Neste modelo, a distribuição estatística do grau de centralidade dos nós da rede tende a seguir uma lei de potência, onde um número muito pequeno de nós concentra muitas ligações (alto *degree centrality*) e um grande número de nós possui pouquíssimas ligações (baixo *degree centrality*). Tal modelo é considerado um modelo *power-law* justamente pelo fato de produzir uma distribuição estatística de graus de centralidade que segue uma lei de potência. As redes que seguem essa distribuição são conhecidas como redes livres de escala (*scale-free networks*).

Alguns modelos matemáticos se mostraram adequados para representação de redes reais. Estudos demonstram que a internet (FALOUTSOS *et al.*, 1999), a *Web* (ADAMIC, 1999), as redes de colaboração científica (NEWMAN, 2001) e as redes de correspondência eletrônica (EBEL *et al.*, 2002) apresentam características de redes do tipo mundo-pequeno, ou seja, baixo diâmetro e alto coeficiente de agrupamento, e são de fato redes livres de escala com a distribuição estatística dos graus de centralidade dos nós seguindo uma lei de potência.

3.2.Métricas de Rede

Diversas são as métricas adotadas por pesquisadores para a abordagem de seus problemas. Existem métricas relacionadas ao tamanho, aos níveis de conectividade e transitividade, ao grau de miscigenação e à estrutura comunitária das redes (NEWMAN, 2003). Não obstante a isso, novas métricas surgem à medida que surgem também novos problemas a serem equacionados. No entanto, existe um conjunto de métricas formalmente estabelecidas e co-

[7] Redes onde cada nó está conectado a um número k fixo de vizinhos.

mumente utilizadas no estudo e resolução de grande parte dos problemas envolvendo redes complexas, sobre as quais cabe uma breve explicação.

Considere um grafo não direcionado $\boldsymbol{G}(\boldsymbol{N},\boldsymbol{L})$, onde $\boldsymbol{N}$ representa o conjunto de nós do grafo, sendo definido por $\boldsymbol{N} = \{\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_g}\}$ e $\boldsymbol{L}$ representa o conjunto de ligações entre pares de nós no grafo, sendo definido por $\boldsymbol{L} = \{\boldsymbol{l_1}, \boldsymbol{l_2}, \ldots, \boldsymbol{l_h}\}$. Considere ainda que $\boldsymbol{g}$ represente o número de nós do grafo e $\boldsymbol{h}$ represente o número de ligações entre pares de nós. Cada ligação em $\boldsymbol{L}$ pode ser representada por:

$$l_k = (n_i, n_j) \Rightarrow \begin{Bmatrix} 1 \le k \le h \\ 1 \le i \le g \\ 1 \le j \le g \end{Bmatrix}$$

Sendo assim, $\boldsymbol{l_k}$ corresponde a uma ligação dedicada a conectar o nó $\boldsymbol{n_i}$ ao $\boldsymbol{n_j}$. Considere também que o número de nós adjacentes (que possuem conexão) a um nó $\boldsymbol{n_i}$ específico é representado por $\boldsymbol{d}(\boldsymbol{n_i})$.

3.2.1. Diâmetro da Rede

O diâmetro da rede ($\boldsymbol{D}$) é uma medida de tamanho que indica a distância geodésica (*geodesic distance*) entre os pares de nós conectados na rede. Em redes com apenas um componente, considerando $\boldsymbol{p_{ij}}$ como a distância geodésica entre $\boldsymbol{i}$ e $\boldsymbol{j}$, o valor do diâmetro pode ser expresso da seguinte forma:

$$D = \frac{1}{g(g+1)/2} \sum_{i \ge j} p_{ij}$$

Geralmente o caminho mínimo (*shortest path*) entre dois nós, ou seja, o menor número de ligações necessárias para sair de um nó e alcançar outro na rede, é considerado como a distância geodésica entre eles. Em redes com mais de um componente, a definição de tal métrica pode se tornar um problema. NEWMAN (2003) apresenta uma solução viável para sua resolução.

3.2.2. Medidas de Centralidade

As medidas de centralidade indicam o grau de conectividade direta entre nós da rede. Dentre as diferentes medidas de centralidade presentes na literatura, destacam-se:

Grau de Centralidade (***DCE***): Número de conexões (de entrada e saída) de cada nó. Equivale ao número de nós adjacentes a um nó e seu valor é dado por $DCE(n_i) = d(n_i)$. Considerando $\boldsymbol{n_{max}}$ como o nó de maior grau de centralidade na rede, a medida normalizada é dada por:

$$NDCE(n_i) = \frac{DCE(n_i)}{DCE(n_{\max})}$$

Grau de Intermediação (***BCE***): Equivale ao número de distâncias geodésicas entre quaisquer dois nós da rede que passam por um nó específico. Tal medida indica o quanto um nó está no caminho mínimo entre outros dois pares de nós. Seja $\boldsymbol{p_{jk}}$ o número de distâncias geodésicas que ligam os nós $\boldsymbol{j}$ e $\boldsymbol{k}$, e $\boldsymbol{p_{jk}(n_i)}$ o número de distâncias geodésicas que passam pelo nó $\boldsymbol{n_i}$. O grau de intermediação é dado por:

$$BCE(n_i) = \sum_{j<k} \frac{p_{jk}(n_i)}{p_{jk}}$$

Considerando $\boldsymbol{n_{max}}$ como o nó de maior grau de intermediação na rede, a medida normalizada é dada por:

$$NBCE(n_i) = \frac{BCE(n_i)}{BCE(n_{\max})}$$

Grau de Proximidade (***CCE***): Inverso da soma das distâncias geodésicas entre cada nó da rede e os demais. Indica o quão próximo um nó da rede está dos demais. Considerando mais uma vez $\boldsymbol{p_{ij}}$ como a distância geodésica entre $\boldsymbol{i}$ e $\boldsymbol{j}$, o valor do grau de proximidade pode ser expresso da seguinte forma:

$$CCE(n_i) = \left[\sum_{j=1}^{g} p_{ij} \right]^{-1}$$

Considerando $\boldsymbol{n_{max}}$ como o nó de maior grau de proximidade na rede, a medida normalizada é dada por:

$$NCCE(n_i) = \frac{CCE(n_i)}{CCE(n_{\max})}$$

3.2.3. Medidas de Transitividade

As medidas de transitividade indicam o grau de conectividade indireta, ou seja, entre vizinho, da rede. Dentre as diferentes medidas de transitividade presentes na literatura, destaca-se:

Coeficiente de Agrupamento ($\boldsymbol{CC1}$): Indica a probabilidade dos vizinhos de um nó da rede se conectarem entre si. Considerando $\boldsymbol{N_i}$ o conjunto de nós vizinhos ao nó $\boldsymbol{i}$, $\boldsymbol{k_i}$ o número de vizinhos de $\boldsymbol{i}$ e $\boldsymbol{c_{jk}}$ = 1 ou 0 caso exista ou não uma ligação entre $\boldsymbol{j}$ e $\boldsymbol{k}$ desde que $\boldsymbol{j}, \boldsymbol{k} \in \boldsymbol{N_i}$, o valor do coeficiente de agrupamento pode ser expresso por:

$$CC1(n_i) = \frac{\sum_{j>k} c_{jk}}{k_i(k_i - 1)/2}$$

Considerando $\boldsymbol{n_{max}}$ como o nó de maior coeficiente de agrupamento na rede, a medida normalizada é dada por:

$$NCC1(n_i) = \frac{CC1(n_i)}{CC1(n_{\max})}$$

4. MEDIDAS DE INFLUÊNCIA

O principal objetivo dos sistemas de recuperação de informação é atender as necessidades de informação dos usuários com desempenho satisfatório, o que pode ser traduzido em oferecer objetos informacionais, tais como documentos, imagens, áudio e vídeos, mais relevantes e, em tempo adequado. Dessa forma, os sistemas de recuperação de informação modelam as necessidades de informação dos usuários sob a forma de consultas que, posteriormente, serão remodeladas dentro de um espaço específico (booleano, vetorial ou probabilístico) e comparadas aos objetos presentes na coleção.

Especificamente no modelo vetorial de recuperação, as consultas e os documentos são representados como vetores de termos, sendo que a similaridade entre eles é dada pelo grau de proximidade entre os vetores, medido pelo coseno do ângulo formado entre eles. Tal medida é utilizada para atribuição do grau de relevância de um documento em relação a uma consulta. Dessa forma, o grau de influência dos termos nos documentos e na coleção exerce papel fundamental no estabelecimento de funções de cálculo de similaridade. Entende-se por grau de influência, ou peso, de um termo, a sua capacidade medida de descrever o conteúdo de um

documento. Em um documento existem termos que carregam maior significado que outros e, por essa razão, são capazes de descrever de maneira mais fidedigna o seu conteúdo.

Diversos pesquisadores propuseram funções para cálculo de similaridade (SALTON; LESK, 1968; SALTON; BUCKLEY, 1988; SINGHAL *et al.*, 1996, ZOBEL; MOFFAT, 1998). No entanto, em sua grande maioria, as funções se baseiam em critérios de freqüência - número de ocorrências do termo no documento - para o estabelecimento do peso de um termo em um documento. Uma instância bastante conhecida desta classe de funções, o $\boldsymbol{TF-IDF}$ (BAEZA-YATES; RIBEIRO-NETO, 1999), estabelece que a similaridade entre as consultas e os documentos pode ser obtida através da função

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \bullet \vec{q}}{\left|\overrightarrow{d_j}\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

onde $\boldsymbol{sim(d_j, q)}$ representa o grau de similaridade entre o documento $\boldsymbol{j}$ e a consulta $\boldsymbol{q}$, $\boldsymbol{w_{i,j}}$ representa o peso do termo $\boldsymbol{i}$ no documento $\boldsymbol{j}$, $\boldsymbol{w_{i,q}}$ representa o peso do termo $\boldsymbol{i}$ na consulta $\boldsymbol{q}$, e $\boldsymbol{t}$ representa o número de termos comuns à consulta e ao documento. Quanto maior o grau de similaridade, mais relevante será considerado o documento em relação à consulta, sendo que esta medida será utilizada para ordenamento de resultados.

O esquema de atribuição de peso aos termos nos documentos descrito pelos autores é dado por $w_{i,j} = f_{i,j} \times idf_i$ onde $\boldsymbol{f_{i,j}}$ é a freqüência do termo $\boldsymbol{i}$ no documento $\boldsymbol{j}$, e $\boldsymbol{idf_i}$ é a freqüência invertida do termo nos documentos da coleção:

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \qquad\qquad idf_i = \log \frac{N}{n_i}$$

Nas funções acima, $\boldsymbol{freq_{i,j}}$ corresponde ao número de ocorrências do termo $\boldsymbol{i}$ no documento $\boldsymbol{j}$, $\boldsymbol{max_l\, freq_{l,j}}$ corresponde ao número de ocorrências do termo $\boldsymbol{l}$ no documento $\boldsymbol{j}$, sendo $\boldsymbol{l}$ o termo que mais vezes ocorreu no documento, $\boldsymbol{N}$ corresponde ao número total de documentos na coleção e $\boldsymbol{n_i}$ corresponde ao número de documentos da coleção em que o termo $\boldsymbol{i}$ ocorre.

Almeida *et al.* (2007) apresentam diversos componentes considerados em diferentes esquemas de atribuição de pesos. Em todos eles a freqüência do termo no documento é o principal fator de influência do termo no documento e na coleção.

5. ABORDAGEM

A abordagem proposta se baseia na representação de documentos como redes complexas de palavras, ou termos, tal qual descrito por Cancho e Solé (2001), Cancho *et al.* (2004), Cancho (2005), Solé *et al.* (2005) e Dorogovtsev e Mendes (2001). Tal forma de representação permite o estabelecimento de relações entre termos possibilitando a extração de métricas de rede a serem utilizadas para atribuição do grau de importância dos termos nos documentos.

De acordo com Solé *et al.* (2005) existem duas formas de construção de redes de termos: partindo das relações semânticas ou das relações sintáticas existentes entre eles. É possível, por exemplo, construir uma rede de termos relacionados sintaticamente a partir da relação de co-ocorrência de pares de termos em frases, ou mesmo a partir da relação de distância entre eles dentro dos documentos. Da mesma forma, é possível construir uma rede de termos relacionados semanticamente a partir da extração e correlação de sintagmas dos documentos ou da extração de relações entre termos de *thesaurus* preconcebidos.

Na presente abordagem são utilizadas relações sintáticas para construção da rede complexa de palavras. Cada documento se constitui em uma rede distinta, onde cada termo se torna um nó e a distância entre eles dentro do documento se torna fator definidor de suas ligações. Mais especificamente, dois termos estão ligados na rede se a distância entre eles no documento for menor ou igual a uma distância máxima predeterminada. O grafo resultante é não direcionado (sem orientação) e não ponderado (com ligações sem valores), o que significa que as relações entre os termos são mútuas - se $\boldsymbol{A}$ está à distância $\boldsymbol{x}$ de $\boldsymbol{B}$, $\boldsymbol{B}$ está à distância $\boldsymbol{x}$ de $\boldsymbol{A}$ - e que múltiplas ocorrências de ligações entre eles não são consideradas. A figura 1 exemplifica o esquema utilizado para representação de um documento como uma rede, em forma de grafo.

Partindo de uma coleção $\boldsymbol{C}$ contendo $\boldsymbol{N}$ documentos, obtém-se uma coleção $\boldsymbol{G}$ de grafos não direcionados, sendo $\boldsymbol{G_j}$ o grafo representativo do documento $\boldsymbol{j}$ da coleção. Para cada grafo da coleção são extraídas as métricas normalizadas de centralidade e de transitividade.

Assim sendo, para cada nó $\boldsymbol{i}$ pertencente ao grafo $\boldsymbol{G_j}$ temos as métricas $\boldsymbol{NDCE_{i,j}}$, $\boldsymbol{NBCE_{i,j}}$, $\boldsymbol{NCCE_{i,j}}$, e $\boldsymbol{NCC1_{i,j}}$ representando, respectivamente, o grau de centralidade, o grau de intermediação, o grau de proximidade e o coeficiente de agrupamento do nó $\boldsymbol{i}$ no grafo $\boldsymbol{G_j}$.

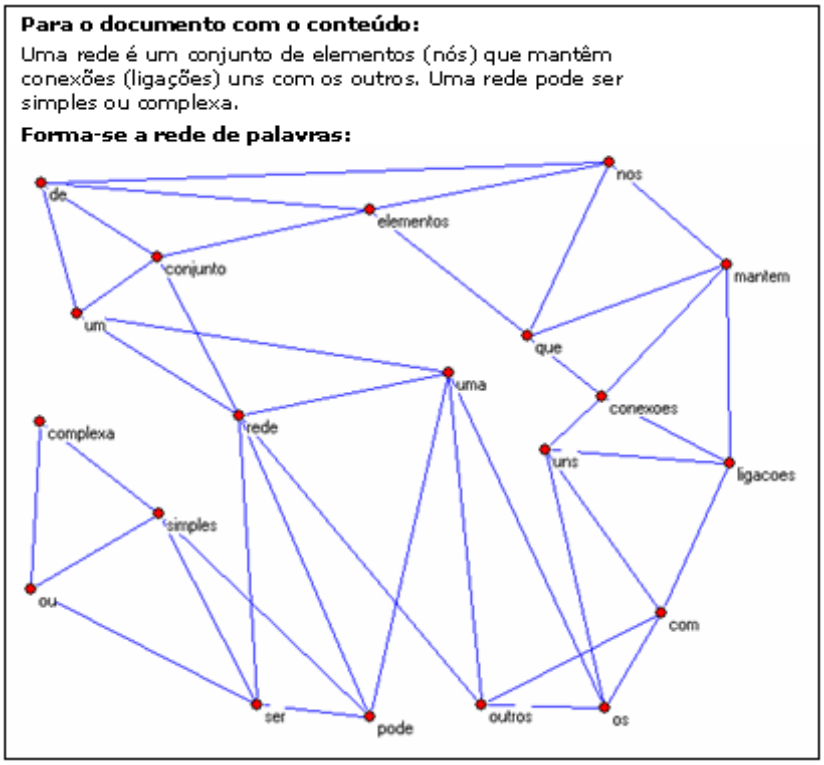

FIGURA 1 - Rede de termos relacionados sintaticamente pelo critério de distância máxima ($\boldsymbol{d}$) com $\boldsymbol{d} = \mathbf{2}$.
Fonte: o autor.

Assim como o $\boldsymbol{TF}$ é utilizado no modelo vetorial para estabelecimento do peso de um termo em um documento ou consulta, tais medidas são utilizadas, uma a uma, dentro da abordagem proposta com o mesmo propósito. Dessa forma, além da função utilizada tradicionalmente no método $\boldsymbol{TF-IDF}$, quatro novas funções serão utilizadas para a atribuição de pesos: Para a função que utiliza a freqüência do termo, denominada $\boldsymbol{TF-IDF}$ temos $w_{i,j} = f_{i,j} \times idf_i$, tal qual definido na sessão 4. Para a função que utiliza o grau de centralidade, denominada $\boldsymbol{NDCE-IDF}$ temos $w_{i,j} = NDCE_{i,j} \times idf_i$. Para a função que utiliza o grau de intermediação, denominada $\boldsymbol{NBCE-IDF}$ temos $w_{i,j} = NBCE_{i,j} \times idf_i$. Para a função que utiliza o grau de proximidade, denominada $\boldsymbol{NCCE-IDF}$ temos $w_{i,j} = NCCE_{i,j} \times idf_i$. E,

finalmente, para a função que utiliza o coeficiente de agrupamento, denominada $\boldsymbol{NCC1-IDF}$ temos $w_{i,j} = NCC1_{i,j} \times idf_i$.

A adoção desta abordagem para cálculo de pesos permitirá avaliar, para uma determinada coleção de documentos e consultas, o quão boa é cada métrica de rede, isoladamente, para substituição da freqüência enquanto indicador de influência dos termos em documentos.

6. EXPERIMENTO E RESULTADOS

Para realização dos experimentos, a coleção CACM (FOX, 1983), composta por 3.204 documentos publicados no *Communications of ACM journal* de 1958 a 1979, foi utilizada. Nela encontram-se disponíveis 52 consultas com uma média de 15 resultados relevantes para cada consulta.

Antes da efetiva representação dos documentos como redes, foi necessário realizar tratamento ao conteúdo dos documentos da coleção. Sinais de acentuação, caracteres especiais e termos compostos de apenas uma letra foram descartados dos documentos. O vocabulário obtido pelo processamento da coleção contém 11.819 termos.

A distância máxima ($\boldsymbol{d}$) considerada para o estabelecimento de ligações entre termos foi $\boldsymbol{d} = \mathbf{2}$. Os 3.204 grafos gerados possuem, em média, 38 nós e 102 ligações. O grafo de menor tamanho possui dois nós e uma ligação entre eles, enquanto que o grafo de maior tamanho possui 249 nós e 737 ligações.

TABELA 1 - Precisão Interpolada em cada nível de revocação resultantes do processamento das 52 consultas utilizando cada uma das 5 funções de atribuição de pesos descritas na sessão 5

Revocação (%)	Precisão Interpolada (%)				
	TF-IDF	NDCE-IDF	NBCE-IDF	NCCE-IDF	NCC1-IDF
0	65,78	57,62	54,04	42,75	31,48
10	51,71	48,23	39,49	35,22	22,46
20	43,11	37,48	30,51	29,67	18,16
30	35,03	32,02	21,85	24,35	14,86
40	29,09	26,23	16,53	21,63	11,48
50	23,42	21,81	12,07	17,94	8,64
60	15,96	15,85	9,16	15,61	6,43
70	11,98	11,43	6,59	9,26	4,59
80	8,75	8,79	5,24	7,30	3,71
90	5,74	5,57	3,69	5,17	3,16
100	5,69	5,54	3,69	4,85	3,14

Fonte: o autor.

Tal como recomendado por Baeza-Yates e Ribeiro-Neto (1999), os resultados apresentados na tabela 1, também foram apresentados em um gráfico, uma vez que a análise gráfica facilita a avaliação da eficiência de um método de recuperação em comparação aos outros. Especificamente nesta abordagem, as variações metodológicas residem apenas na mudança da função de atribuição de pesos. Sendo assim, a análise do gráfico 1 permite avaliar o nível de eficiência de cada função de atribuição proposta.

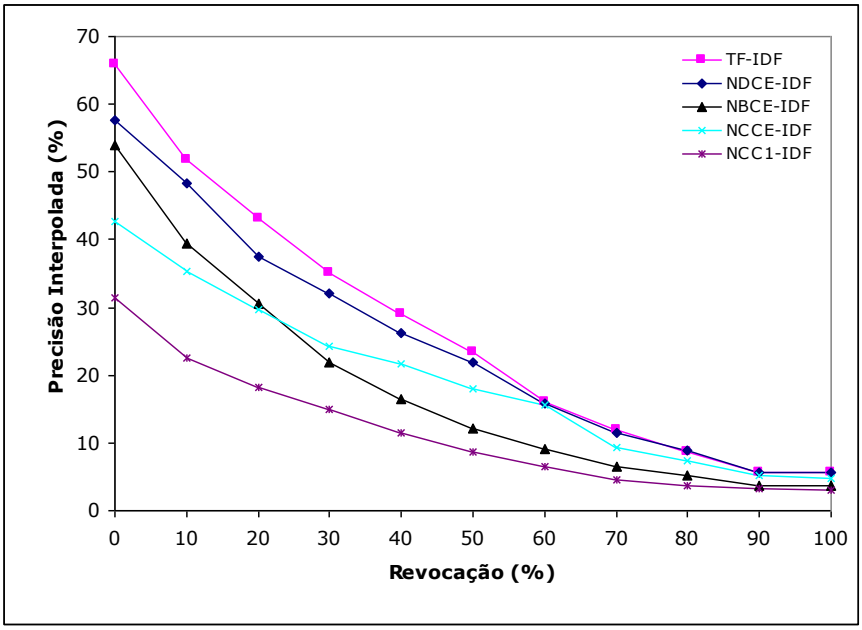

GRÁFICO 1 - Precisão Interpolada versus Revocação resultantes do processamento das 52 consultas utilizando cada uma das 5 funções de atribuição de pesos descritas na sessão 5.
Fonte: o autor.

No gráfico 1, é possível observar que a abordagem que apresentou melhores resultados foi a que utilizou o $\boldsymbol{TF-IDF}$ como método para atribuição de pesos a termos em documentos, seguida de $\boldsymbol{NDCE-IDF}$. A que apresentou piores resultados foi a $\boldsymbol{NCC1-IDF}$. As abordagens $\boldsymbol{NBCE-IDF}$ e $\boldsymbol{NCCE-IDF}$ apresentaram resultados intermediários. Cabe salientar que a abordagem $\boldsymbol{NBCE-IDF}$ se saiu melhor a uma revocação de até 24%, apresentando taxas de precisão melhores nesta faixa. No entanto, os resultados pioraram significativamente após esta faixa, tendo se aproximado da pior abordagem ($\boldsymbol{NCC1-IDF}$). Já a abordagem $\boldsymbol{NCCE-IDF}$, apesar de apresentar taxas de precisão piores que a $\boldsymbol{NBCE-IDF}$ a uma revocação inferior a 24%, se aproximou dos resultados de $\boldsymbol{TF-IDF}$ e $\boldsymbol{NDCE-IDF}$ a taxas de revocação maiores.Importante também destacar que as abordagens $\boldsymbol{TF-IDF}$ e $\boldsymbol{NDCE-IDF}$ se equivalem a taxas de revocação superiores a 60%.

7. Conclusão

Mesmo refutando-se a hipótese de que o uso de algumas métricas de redes complexas, isoladamente, pudesse substituir com maior eficiência a medida de freqüência nas fórmulas de

atribuição de pesos a termos em documentos do esquema $\boldsymbol{TF-IDF}$, existem evidências de que uma composição entre tais métricas possa proporcionar melhores resultados. Aparentemente, termos com maior grau de influência apresentam, em média, graus de proximidade mais elevados e baixos coeficientes de agrupamento. Além disso, o mecanismo para atribuição de pesos utilizado no modelo vetorial, que considera a freqüência invertida do termo ($\boldsymbol{IDF}$), pode ser inapropriado para ser usado em conjunto com métricas de redes complexas, o que demanda pesquisa na construção e avaliação de novos mecanismos.

Destaca-se ainda a necessidade de pesquisas no campo da identificação e caracterização de elementos lingüísticos em redes de termos a fim de se identificar padrões de relacionamentos sintáticos e semânticos para, a partir daí, utilizar tais padrões para correlacionar métricas de rede. Cabe ressaltar que não foram avaliadas relações semânticas e outros tipos de relações sintáticas entre termos. Variações nas distâncias máximas utilizadas para o estabelecimento de relações sintáticas entre termos também podem ser consideradas.

A construção de um novo modelo de recuperação baseado num espaço vetorial multidimensional onde consultas e documentos possam ser representados como grafos e, funções de cálculos de similaridades entre grafos possam ser utilizadas, se apresenta como um caminho a ser explorado.

Abstract: *In the last years, the information retrieval field has received much attention from the world scientific community. Research on the improvement of methods and algorithms for textual information retrieval has increased, largely concentrated in the improvement of vector model, especially in efficient methods and functions for similarity calculation between documents and queries. In parallel, the networks analysis subject has attracted the interest of the scientific community due to its ability to represent complex issues in an objective manner, offering a theoretical and practical approach for the study of the properties and behavior of the elements and relations of which problems are made. Recently, research papers considering documents as word complex networks has been developed. However, using this approach to solve information retrieval and classification problems has been under-exploited. This paper presents an approach, based on metrics of complex networks, that obtain functions to assign weights to terms in documents. The approach performs as well as a vector model based approach, when applied to estimate the similarity between documents and queries from a reference collection. This demonstrates the applicability of the metrics of word complex networks in information retrieval problems.*
Keywords: *Information retrieval, complex networks, word networks, network metrics.*

REFERÊNCIAS

ADAMIC, L. A. The Small World Web. In: EUROPEAN CONFERENCE ON RESEARCH AND ADVANVED TECHNOLOGY FOR DIGITAL LIBRARIES, 3., 1999, Paris. **Anais…**London: Springer-Verlag, 1999. p. 443-452.

ALMEIDA, H. M.; GONÇALVES, M. A.; CRISTO, M.; CALADO, P. A combined component approach for finding collection-adapted ranking functions based on genetic programming. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 30., 2007, Amsterdam. **Anais…**New York: ACM, 2007. p. 399-406.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. 1. ed. New York: Addison Wesley, 1999. 544 p.

BARABÁSI, A.; ALBERT, R. Emergence of scaling in random networks. **Science**, New York, v. 286, n. 5439, p. 509-512, out. 1999.

BONACICH, P. Factoring and weighting approaches to status scores and clique identification. **Journal of Mathematical Sociology**, Philadelphia, v. 2, n. 1, p. 113-120, 1972.

BRANDE, U.; HOEFER, M.; LERNER, J. WordSpace - Visual Summary of Text Corpora. In: INTERNATIONAL SYMPOSIUM ON ELETRONIC IMAGING, 18., 2006, San Jose. **Anais…**[S. I. : s. n.], 2006. p. 212-223.

BRANDES, U.; CORNELSEN, S. Visual Ranking of Link Structures. In: INTERNATIONAL WORKSHOP ON ALGORITHMS AND DATA STRUCTURE, 7., 2001, Providence. **Anais…**London: Springer-Verlag, Brown University, 2001. 11 p.

BURGESS, M.; CANRIGHT G.; ENGØ-MONSEN, K. **Importance-ranking functions derived from the eigenvectors of directed graphs**. [S. I. : s. n.], 2006, 38 p. (DELIS Technical Report DL-TR-0325).

CANCHO, R. F.; SOLÉ, R. V. The Small World of Human Language. **The Royal Society B**, London, v. 268, p. 2261-2266, 2001.

CANCHO, R. F.; SOLÉ, R. V.; KÖHLER, R. Patterns in syntactic dependency networks. **Physical Review**, [S. I.], v. 69, p. 8, 2004.

CANCHO, R. F. The structure of syntactic dependency networks: insights from recent advances in network theory. In: ALTMMAN, G.; LEVICKIJ, V.; PEREBYINIS, V. **The Problems of quantitative linguistics**, Chernivtsi: Ruta, 2005. p.60-75.

CHITRAPURA, K. P.; KASHYAP, S. R. Node Ranking In Labeled Directed Graphs. In: ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 13., 2004, Washington. **Anais…**NewYork: ACM, 2004. p. 597-606.

DOROGOVTSEV, S. N.; MENDES, J. F. F. Language as an Evolving Word Web. **The Royal Society B**, London, v. 268, p. 2603-2606, 2001.

EBEL, H.; MIELSCH, L.; BORNHOLDT, S. Scale-free topology of e-mail networks. **Physical Review**, [s. n.], v. 66, n.3, p. 4, 2002.

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

_____. On the evolution of random graphs. **Publications of the Mathematical Institute of the Hungarian Academy of Sciences**, [s. n.], v. 5, p. 17-61, 1960.

ERKAN, G.; RADEV, D. R. LexPageRank: Prestige in Multi-Document Text Summarization. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2004, BARCELONA. **Anais…**[S. I. : s. n.], 2004. p. 365-371.

_____. LexRank: Graph-based lexical centrality as salience in text summarization. **Journal of Artificial Intelligence Research**, AI Access Foundation, v. 22, n. 1, p. 457-479, 2004.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the Internet topology. In: CONFERENCE ON APPLICATIONS, TECHNOLOGIES, ARCHITECTURES, AND PROTOCOLS FOR COMPUTER COMMUNICATION, 1999, Cambridge. **Anais…**New York: ACM, 1999. p. 251-262.

FOX, E. **Characterization of two new experimental collections in computer and information science containing textual and bibliographical concepts**. Cornell: Cornell University, 1983, 67 p. (Technical Report TR83-561).

GUARE, J. **Six Degrees of Separation**: A Play. New York: Vintage, 1990. 73 p.

HUANG, X.; LAI, Wei. NodeRank: A New Structure Based Approach to Information Filtering. In: INTERNATIONAL CONFERENCE ON INTERNET COMPUTING, 2003, Las Vegas. **Anais…**[S. I.]: CSREA Press, 2003. p. 167-173.

KLEINBERG, J. M. et al. The Web as a graph: Measurements, models and methods. In: INTERNATIONAL CONFERENCE ON COMBINATORICS AND COMPUTING, 5., 1999, Tokyo. **Anais…**Berlin: Springer, 1999. p. 1-17.

KLEINBERG, J. M. Navigation in a small world. **Nature**, [S. I.], v. 406, p. 845, 2000.

_____. The small-world phenomenon: An algorithmic perspective. In: ACM SYMPOSIUM ON THEORY OF COMPUTING, 32., 2000, Portland. **Anais…**New York: ACM, 2000, p. 163-170.

KORFIATIS, N.; SICILIA, M.; HESS, C.; STEIN, K.; SCHLIEDER, C. Social Network Models for Enhancing Reference Based Search Engine Rankings. In: GOH, D.; FOO, S. **Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively**, 1. ed., [S. I.]: Idea Group Inc., 2007, p. 87-107.

KURLAND, O.; LEE, L. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 28., 2005, Salvador. **Anais…**New York: ACM, 2005. p. 306-313.

MONTES-Y-GÓMES, M.; LÓPEZ-LÓPEZ, A.; GELBUKH, A. Information Retrieval with Conceptual Graph Matching. In: INTERNATIONAL CONFERENCE ON DATABASE

AND EXPERT SYSTEMS APPLICATIONS, 11., 2000, London. **Anais…**London: Springer-Verlag, 2000. p. 312-321.

NEWMAN, M. E. J. The structure of scientific collaboration networks. In: NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 2001. **Proceedings…**[S. I. : s. n.], 2001. v. 98, n. 2. p. 404-409.

_____. The structure and function of complex networks. In: SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS, 2003. **Proceedings…**[S. I. : s. n.], 2003. v. 45, n. 2. p. 167–256.

SALTON, G.; LESK, M. Computer evaluation of indexing and text processing. **Journal of the ACM**, ACM, v. 15, n. 1, p. 8-36, 1968.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic retrieval. **Information Processing and Management**, Elsevier, v. 24, n. 5, p. 513-523, 1988.

SINGHAL, A.; BUCKLEY, C.; MITRA, M. Pivoted Document Length Normalization. In: In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 19., 1996, Zurich. **Anais…**New York: ACM, 1996. p. 21-29.

SOLÉ, R. V. et al. Language Networks: their structure, functions and evolution. **Trends in Cognitive Sciences**, [s. n.], 2005.

SRIHARI, R. K. et al. Contextual Information Retrieval using Concept Chain Graphs. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 2005, Paris. **Anais…**[S. I. : s. n.], 2005. p. 1-12.

WATTS, D.; STROGATZ, S. Collective dynamics of ‘small-world’ networks. **Nature**, [S. I.], v. 393, p. 440-442, 1998.

XIE, Z. Centrality Measures in Text Mining: Prediction of Noun Phrases that Appear in Abstracts. In: ACL STUDENT RESEARCH WORKSHOP, 2005, Ann Arbor. **Anais…**Morristown: ACL, 2005. p. 103-108.

ZHOU, X. et al. Relation-based Document Retrieval for Biomedical Literature Databases. In: INTERNATIONAL CONFERENCE ON DATABASE SYSTEMS FOR ADVANCED APPLICATIONS, 11., 2006, Singapore. **Anais…**[S. I. : s. n.], 2006. p. 689-701.

ZOBEL, J.; MOFFAT, A. Exploring the Similarity Space. **ACM SIGIR Forum**, ACM, v. 32, n. 1, p. 18-34, 1998.