UMA REDE SOCIAL CONSTRUÍDA A PARTIR DE DOCUMENTOS DIGITAIS DO PORTAL DA UNIVERSIDADE FEDERAL DE TOCANTINS

A Social Network Built from Digital Documents from the UFT Website

Gentil Barbosa (1), David Nadler Prata (2), Rogério Nogueira de Sousa (3), Rafael Murta (4), Elencarlos Soares Silva (5)

(1) Universidade Federal do Tocantins, Brasil, gentil@uft.edu.br

(2) ddnprata@uft.edu.br (3) rogerio@uft.edu.br (4) rafael.mansilha@gmail.com

(5) elencarlos@uft.edu.br

Resumo

Esta pesquisa propõem a exemplificação de um mapeamento de uma rede social. Desata forma, nesta pesquisa, reproduziu-se uma rede social de entes relacionados à Universidade Federal do Tocantins (UFT) por meio do uso de 13 mil documentos digitais. Para mapear as conexões dos documentos e gerar o grafo, foi utilizado como critério que caso o nome de duas pessoas estejam em um mesmo documento eles possuem uma conexão, para cada arquivo diferente que houver conexão, esta relação é fortalecida. O grafo apresentou 114.405 vértices, e 21.081.984 arestas. Dos dez nós de maior centralidade, os entes encontrados ocuparam, ou ainda ocupam as seguintes funções: reitor, diretor de campus, pró-reitor e vice-reitor. Dessa forma, considera-se a reprodução de uma rede complexa de relacionamentos, com a utilização de documentos digitais como uma alternativa viável para o mapeamento de interações sociais de núcleos que provavelmente não estão mapeados pelas redes sociais online convencionais. Assim, o grafo criado por esse modelo de rede social, construída a partir de documentos digitais de texto, apresenta uma alternativa para mapear relações entre entes, podendo ter diversas finalidades.

Palavras-chave: Redes Sociais; Redes Complexas; Analise de Documentos; Grafo.

Abstract

This research proposes the exemplification of a mapping of a social network. Thus, in this research, a social network of entities related to the Federal University of Tocantins (UFT) was reproduced using 13 thousand digital documents. To map the connections of the documents and generate the graph, it was used as a criterion that if the name of two persons are in the same document they have a connection, for each different file that has a connection, this relationship is strengthened. The graph had 114,405 vertices, and 21,081,984

edges. Of the ten most central nodes, the entities found occupied, or still occupy, the following functions: dean, campus director, pro-rector and vice-rector. In this way, the reproduction of a complex network of relationships is considered, with the use of digital documents as a viable alternative for the mapping of social interactions of nuclei that are probably not mapped by conventional online social networks. Thus, the graph created by this social network model, built from digital text documents, presents an alternative to map relationships between entities, which can have different purposes.

Keywords: Social Networks; Complex Networks; Document Analysis; Graph.

1 Introdução

A Universidade Federal do Tocantins (UFT) é composta por: servidores técnico-administrativos, servidores docentes que juntos totalizam 2178 indivíduos ativos Presidência da República (2021), e discentes, que são aproximadamente 15 mil com vínculo ativo só em graduação. Esses indivíduos são dos mais diversos grupos econômicos, sociais e étnicos. A universidade denomina tais grupos como comunidade da UFT.

Essa comunidade, com um ambiente de entes diversificados, é uma rede de relacionamentos que vão além da relação entre aluno e professor, uma vez que, conforme o artigo 207 da Constituição Federal BRASIL (1988), as Universidades Federais devem obedecer ao princípio da indissociabilidade entre ensino, pesquisa e extensão. Considerando que estes relacionamentos são construídos no interesse acadêmico, ou profissional, a universidade não possui um mapeamento deles.

Durante a passagem de indivíduos pela UFT, é comum que ocorra o registro de seus nomes em documentos, desta forma, esta pesquisa tem por objetivo realizar o mapeamento dos referidos relacionamentos da instituição com base em documentos digitais de texto (txt, pdf, doc, docx, odt, etc.), armazenados pela instituição e quais as finalidades em que este modelo de construção de redes implica.

1.1 Redes de Relacionamento

Para Marteleto (2018), as redes de relacionamentos, ou Redes Sociais, dentro do campo das Ciências Sociais, não se limitam a uma área de estudo em específico, pois podem ser encontradas entre áreas como Antropologia, Sociologia, Economia, áreas da Tecnologia, entre outras. Entretanto, o autor define o conceito de redes sociais como uma forma de compreensão da

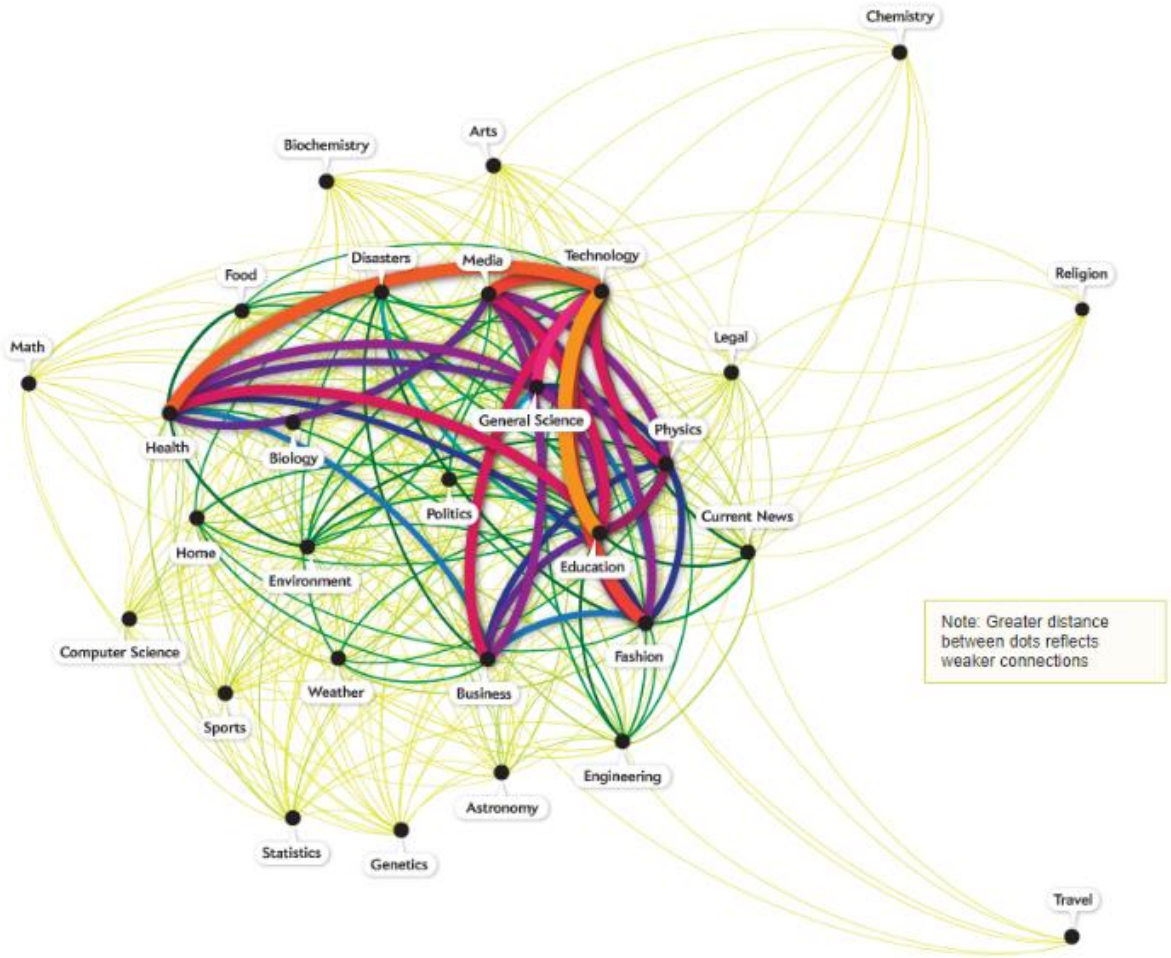
sociedade apoiada nos vínculos relacionais entre indivíduos, que podem levar ao reforço de capacidades de atuação, compartilhamento de conhecimento, captação de recursos e mobilização.

Recuero (2011) define que dois fatores são necessários para uma rede social: o primeiro fator são os atores, que constituem os nós da rede, e é composto por pessoas, instituições ou grupos; já o segundo fator é formado pelas conexões dos atores, podendo ser interações ou laços sociais, considerando assim as redes sociais como uma representação para analisar padrões de conexão de grupos sociais com base nas conexões entre os atores. Em seu estudo sobre as redes de relacionamentos na rede mundial de computadores, Recuero (2011) considera que os indivíduos e suas interações podem ser representados, de forma metafórica, matematicamente, por grafo, a representação de uma rede formada por nós (indivíduos) e arestas (suas relações).

1.2 Redes Complexas e Redes de Relacionamento

Em seu trabalho publicado na revista Scientific American, Fischetti (2011) apresentou um infográfico interativo, ilustrado na figura 1, onde um cientista da BitLy, site de encurtamento de endereço *URL*, examinou 600 páginas e rastreou 6.000 páginas que as pessoas visitaram após acessarem uma dessas 600 páginas.

Figura 1 – Infográfico de Tráfego na Rede de Amantes da Ciência



Fonte: Fischetti (2011)

Esse experimento construiu uma rede complexa associando os conteúdos das páginas da internet acessadas pelas pessoas. Como resultado, observou-se que havia algumas associações inesperadas, como, por exemplo, pessoas que estão interessadas em física também estão interessadas em ciência da computação, porém estas mesmas pessoas também estão muito interessadas em moda.

Inúmeros problemas do mundo real podem ser reproduzidos através de redes complexas, nas quais se conectam os nós. Existem diversos tipos de redes, como redes de cadeia de DNA, redes elétricas, redes aéreas, redes sociais (*online*), entre outras. No campo dos estudos das redes sociais, medidas de influência social vêm sendo elaboradas Newman (2005), Borgatti & Everett, A graph-theoretic perspective on centrality (2006), Borgatti, Identifying sets of key players in a

social network (2006), Agneessens, Borgatti, & Everett (2017). Em grafos como modelos para redes sociais, nós importantes são considerados mais centrais na rede Wasserman, Faust, & others (1994).

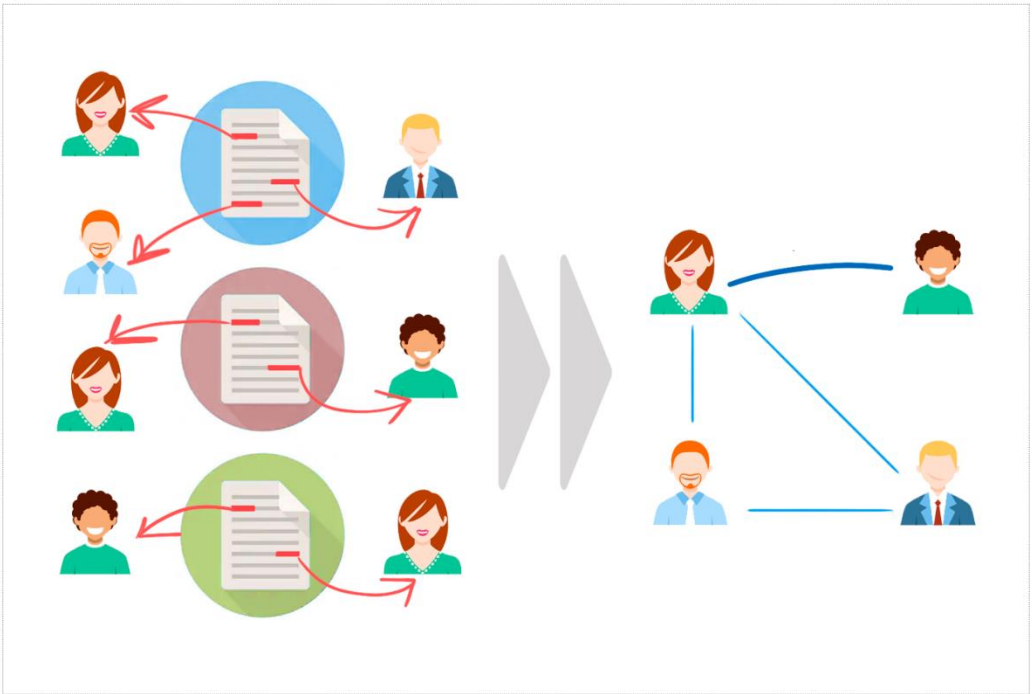
Uma rede complexa pode ser denotada matematicamente por meio de grafos. O grafo (G) é formado por dois conjuntos, sendo um de vértices, $V(G)$, que representam objetos, e outro de arestas, $E(G)$, que correspondem à relação entre os vértices Coppin (2017). As arestas podem ser identificadas por uma tupla (i,j). Em teoria dos grafos, V define a ordem do grafo ao mesmo tempo em que o número de arestas E define o seu tamanho. Diversas propriedades e características podem ser analisadas em um grafo, e as propriedades estudadas são importantes para entendimento das redes complexas.

2 Metodologia

O uso do conceito de redes complexas para a construção de redes com bases de dados mostra-se uma excelente ferramenta para encontrar associações e realizar análises. Considerando que a UFT dispõe de vários tipos de repositórios para armazenamento de documentos digitais, optou-se por utilizar a base de dados de 13 mil documentos digitais, disponíveis de forma pública no portal da UFT:. www.uft.edu.br

Para a pesquisa aqui explicitada, foi considerado que, se duas ou mais pessoas estão em um mesmo documento digital de texto, então elas possuem uma relação. Esses indivíduos podem até não se conhecer de fato, mas definitivamente há uma correlação de interesse entre eles. Sendo assim, os vértices são as pessoas identificadas durante a análise dos documentos que compõem o corpus, e as arestas são formadas a partir da ocorrência de duas ou mais pessoas no mesmo documento; este processo é ilustrado na figura 2. Logo, se os nomes de duas pessoas estão no mesmo documento, elas estão diretamente conectadas.

Figura 2 – Processo de criação da rede complexa de relacionamentos dos entes da UFT

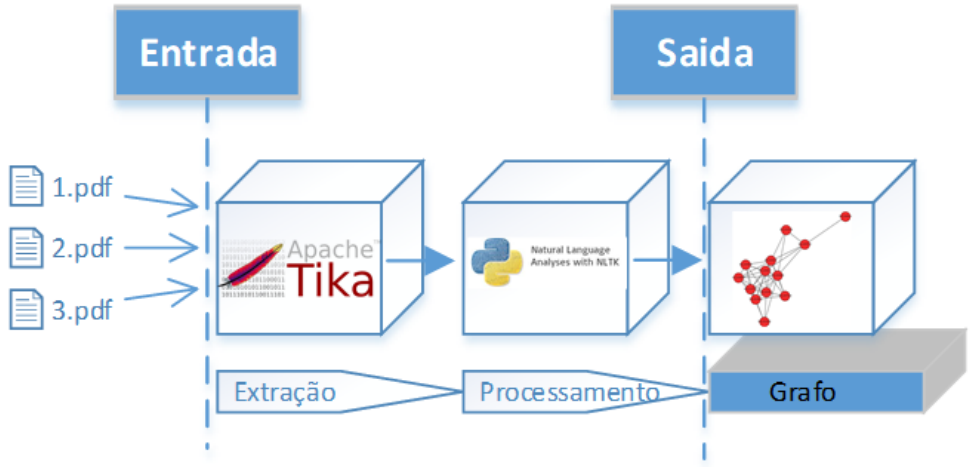


Fonte: Elaborada pelo autor

Os 13 mil documentos utilizados neste experimento foram extraídos do portal da UFT, sendo essa extração feita pela Coordenação de Desenvolvimento de Software da UFT, por meio de uma ferramenta própria de exportação da plataforma utilizada para disponibilizar o Portal. Para o experimento, foram coletados apenas documentos disponíveis de forma pública no site. Para fins de replicação deste estudo, pode-se utilizar um *web crawler*(1) para coletar os arquivos.

Neste acervo de documentos, há apenas documentos que são disponíveis para *download*, e não foram coletadas publicações ou artigos do portal. Em sua grande maioria, foram coletados documentos com as extensões doc, docx e pdf. Ao caracterizar os documentos selecionados para o experimento, os principais tipos de documentos encontrados foram normativas, resoluções, editais, manuais, instruções, relatórios, informes e panfletos. Além disso, conforme representado pela figura 3, o Apache Tika(2) foi utilizado para extrair o texto dos documentos.

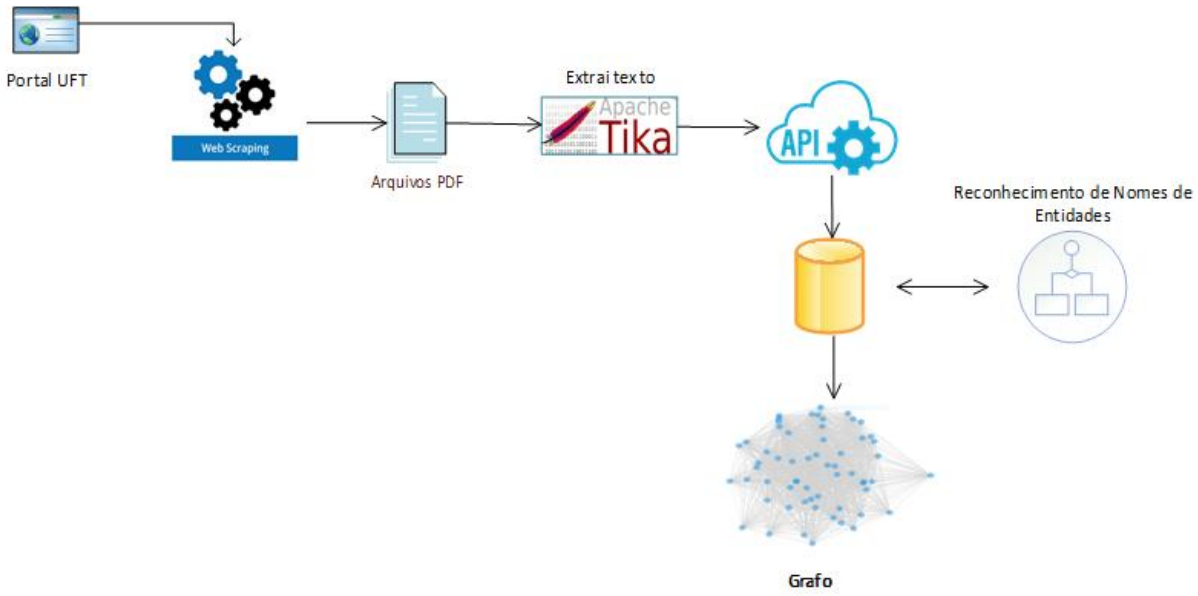
Figura 3 – Processo de extração de nome para criação da rede complexa de relacionamentos dos entes da UFT



Fonte: Elaborada pelo autor

Para formação dos vértices foi utilizada a técnica de Reconhecimento de Entidades Nomeadas (REN), que se refere à tarefa de identificação e classificação de unidades de informação capaz de referenciar entidades, como nomes de pessoas, organizações, locais e datas, a partir de fontes de dados não estruturados como documentos, tarefa essa largamente utilizada no campo do Processamento de Linguagem Natural (PLN) (Nadeau & Sekine, 2007). Todo este processo é representado pela figura 4.

Figura 4 – Formação do Grafo



Fonte: Elaborada pelo autor

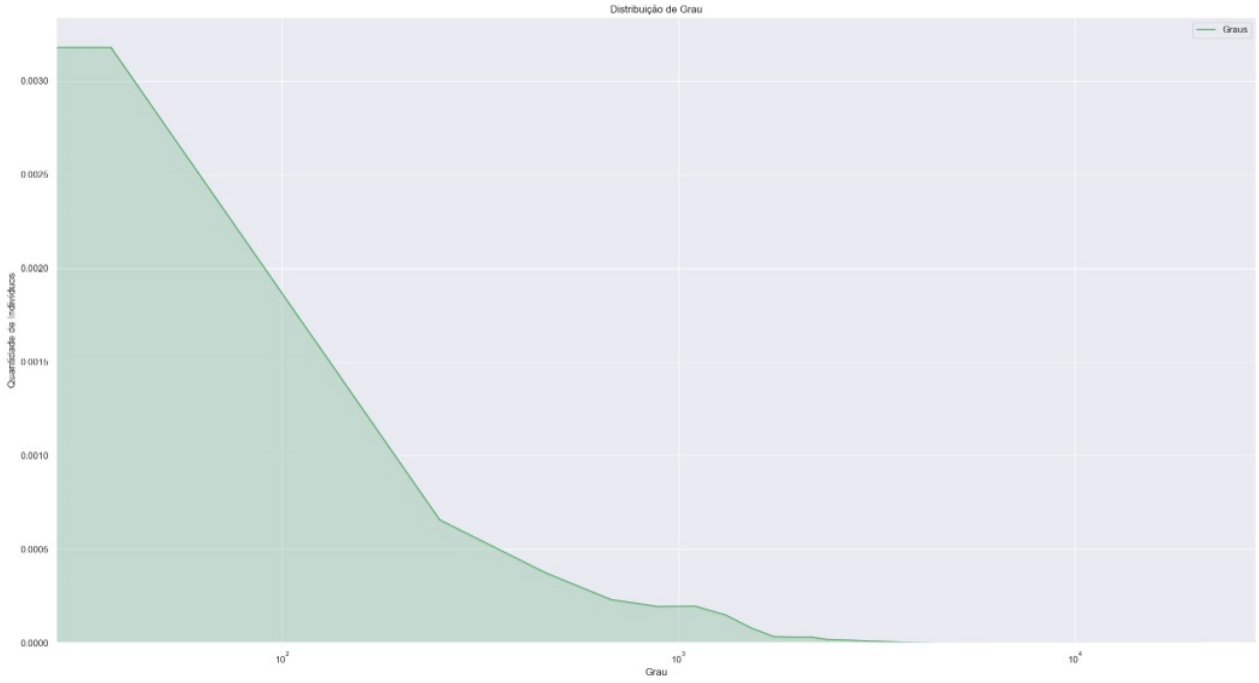
Devido à semelhança entre os documentos extraídos do portal UFT e textos jurídicos, optamos pela implementação das técnicas de REN fazendo uso do *dataset Named Entity Recognition in Brazilian Legal Text* (LeNER-BR[3]) em conjunto com o modelo de arquitetura *Long Short-Term Memory - Conditional Random Fields* (LSTM-CRF) (Araujo, et al., 2018), desenvolvido para recolhimento de entidades nomeadas contidas em textos legais em língua portuguesa. O LeNER-BR possibilita a identificação de entidades dos seguintes tipos: Pessoa, Organização, Local, Tempo, Legislação e Jurisprudência. Para este trabalho, utilizou-se apenas as palavras classificadas como Pessoa, para formação dos vértices do grafo. Cabe salientar que o LSTM-CRF (Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016) atingiu o índice de precisão de 91.80% no que tange à classificação de entidades do tipo Pessoa ao ser treinado e testado no Corpus Paramopama (Mendonça, et al., 2015).

O peso das arestas é dado pela quantidade de documentos em que o mesmo par de indivíduos são citados, portanto, quanto mais documentos citam um par de pessoas distintas, mais conectados eles são. Dessa forma, o experimento trata da análise de uma rede complexa do tipo ponderada.

3 Resultados e Discussões

A distribuição de graus é relevante para a caracterização de uma rede. Nesse sentido, podemos observar no gráfico de distribuição de grau plotado com escala logarítmica, representado na Figura 5, a presença um pequeno grupo de indivíduos que possuem uma grande quantidade de conexão; em contrapartida, um grande grupo de indivíduos apresenta uma pequena quantidade de conexões. Os vértices do pequeno grupo com alta conectividade são denominados *hubs*[4] e ficam ao longo da cauda apresentada no gráfico da figura 5.

Figura 5 – Distribuição de Grau



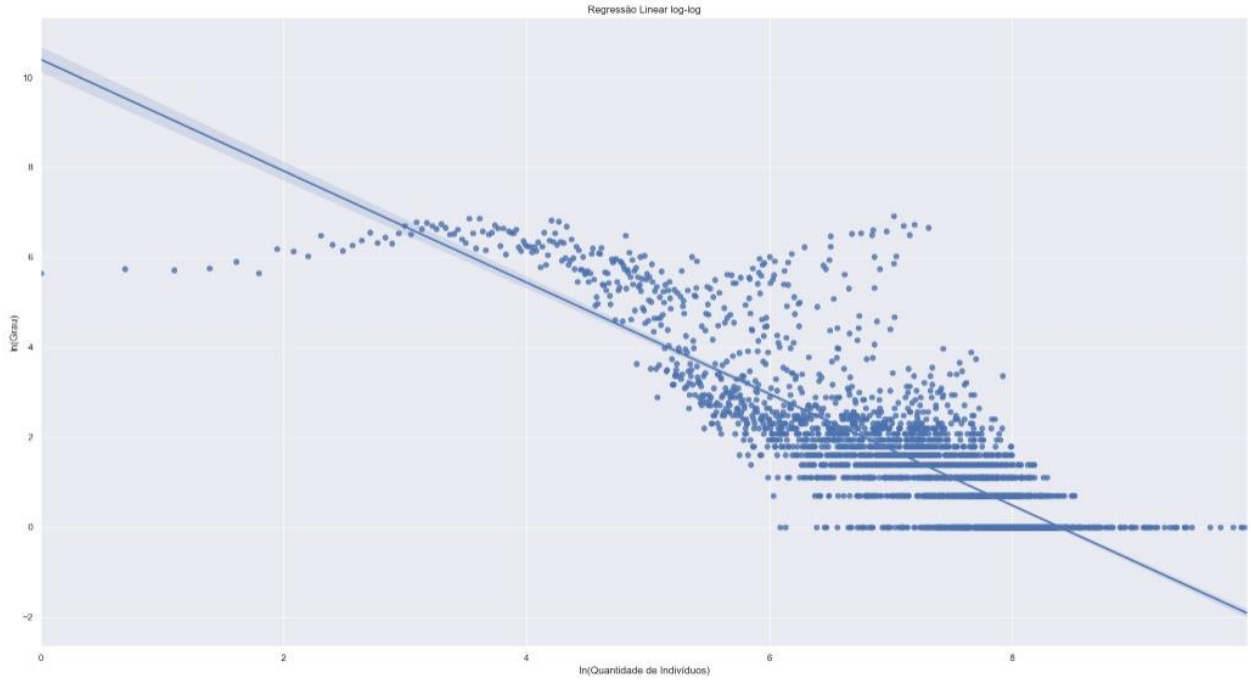
Fonte: Elaborada pelo autor

Observou-se que, ao analisar os nós altamente conectados, os 10 principais entes encontrados foram, ou são egressos e discentes que participaram de diversos editais, como processos seletivos para ingresso à universidade, auxílios estudantis, seleção de bolsas, Exame Nacional de Desempenho dos Estudantes, e outros do gênero.

Na figura 6 pode-se observar de forma mais explícita a existência dos *hubs*, e que a dispersão das frequências de grau acompanha a linha da regressão aplicada à distribuição na escala

log-log, demonstrando a relação matemática entre os escalares frequência e grau. Logo, a rede aqui estudada tende a ser uma rede de livre escala Clauset, Shalizi, & Newman (2009).

Figura 6 – Regressão linear logarítmica



Fonte: Elaborada pelo autor

O grafo gerado apresentou o número de 114.405 vértices, ou entes, com 21.081.984 arestas. O grau médio foi de 368,55, este valor alto se dá a alta conectividade dos nós do grafo. Levando em consideração que os nós são entes, em média, uma pessoa se conecta com outras 368 indivíduos, e por ser uma média alta, dado o contexto, isso implica que poucas pessoas não estão conectadas à rede.

Já a média de clusterização é de aproximadamente 0,909, isso quer dizer que, para cada nó, há aproximadamente 90% de probabilidade de um nó X conectado ao nó Y, também possuir uma aresta com outro nó Z que também está conectado ao nó Y.

A tabela 1 apresenta uma lista dos 10 nós de maior centralidade de grau dentro da rede complexa social gerada. Dos nomes apresentados, três entes ocuparam, ou ocupam, o cargo de reitores da UFT, outros três ocupam o cargo de diretor de câmpus; nove ocupam, ou ocuparam, alguma pró-reitoria e três ocuparam, ou ocupam, o cargo de vice-reitor.

Outra forma de aferir a importância de um vértice para uma rede complexa é medir sua centralidade autovetor Bonacich (1987), o que, assim como a centralidade de grau, leva em consideração a quantidade de nós diretamente relacionados. Essa medida também considera as ligações com nós altamente conectados, o que significa que, além do seu grau de conexão, ela calcula também a importância dos seus vizinhos.

Tabela 1- Lista dos 10 nós com maior centralidade de grau

Entidade	**Cargo/Relevância**	**Centralidade**	**Grau**
Indivíduo número 1	Reitor(a)	0.1811	20722
Indivíduo número 2	Pró-reitor(a)	0.1741	19923
Indivíduo número 3	Vice-Reitor	0.1712	19589
Indivíduo número 4	Reitor(a)	0.1608	18407
Indivíduo número 5	Pró-reitor(a)	0.1447	16565
Indivíduo número 6	Reitor(a)	0.1324	15154
Indivíduo número 7	Pró-reitor(a)	0.1139	13040
Indivíduo número 8	Pró-reitor(a)	0.1090	12474
Indivíduo número 9	Pró-reitor(a)	0.1077	12328
Indivíduo número 10	Pró-reitor(a)	0.1076	12312

Fonte: Dados da pesquisa

O repositório utilizado para o armazenamento dos documentos do portal, conforme servidor responsável, lotado na coordenadoria de Infraestrutura de Tecnologia da Informação da Universidade, foi implementado entre os anos de 2011 e 2012, nos últimos anos em que o Professor Alan Barbiero ocupou o cargo de reitor da UFT. Desta forma, o Grafo que resultou na rede social é um reflexo da gestão entre 2012 e 2020.

O mapeamento apresentou várias relações de professores de áreas de ensino diferente, como Ciência da Computação, Letras e Ciências Contábeis, que compunham um comitê de governança digital, mas que não usavam redes sociais, ou não possuíam estas conexões mapeadas.

Observou-se também a possibilidade de mapear grupos de trabalho da universidade de acordo com a área de atuação dos professores, e com isso detectar as possíveis conexões inexistentes, ou fracas, e definir no plano estratégico da universidade ações voltadas para melhorar a interdisciplinaridade dentro da universidade, contudo como as informações sobre as pessoas,

seus cargos na universidade, são dados que não possuem acesso público, não foi possível realizar esta análise para esta pesquisa.

Dentro das análises levantadas, mas que também estão limitadas quanto ao acesso de dados não públicos, o uso do mapeamento para encontrar pessoas com conexões fracas, ou até mesmo inexistentes, com o propósito de classificar estas ocorrências em algo que é característico das atribuições do servidor, ou se há uma questão que ocasiona esta ocorrência e assim aprimorar as relações das pessoas quem compõem a universidade.

4 Considerações Finais

Dada a relevância da indissociabilidade entre ensino, pesquisa e extensão que as universidades devem observar, conforme definido na Constituição Federal BRASIL (1988), e considerando o propósito da extensão como o desenvolvimento de ações que busquem provocar a troca de conhecimentos, podemos observar que a UFT possui relacionamentos com pessoas que vão além da sua própria comunidade, pois a universidade se relaciona com indivíduos que não estão necessariamente dentro da universidade.

Sendo assim, dentro dos 114.405 entes encontrados na rede social complexa formada por este estudo, há uma quantidade expressiva de pessoas externas à comunidade da UFT. Contudo, uma limitação desta pesquisa é que não foi possível obter acesso a todos os nomes de indivíduos desta comunidade, uma vez que estes dados não são públicos, para classificar os entes encontrados e apresentar uma análise das relações do público externo e suas relações para com a comunidade da UFT.

Tendo em vista que, se dois ou mais indivíduos estão em um mesmo documento, a rede complexa aqui construída define que há uma relação entre eles, mesmo que estas pessoas não se conheçam pessoalmente (o que evidencia uma relação de interesse mútuo entre tais indivíduos), o uso deste tipo de ferramenta pode inclusive ajudar na criação de novos relacionamentos entre indivíduos que possuem interesses semelhantes.

Assim, considerando que as redes sociais online disponíveis através da internet possuem a limitação de que as pessoas, assim como suas relações para com outras pessoas, são quem definem

estas conexões, uma vez que sua adesão é voluntária, a reprodução de uma rede complexa de relacionamentos, com a utilização de documentos digitais, mostrou-se uma alternativa viável para o mapeamento de redes sociais, permitindo o estudo de entes que, por sua vez, podem não estar mapeados nas redes sociais *online* convencionais, como, por exemplo, dois colegas de trabalho que não são amigos, e por consequência não criaram um vínculo em redes sociais online, mas possuem um convívio devido a relação profissional. Sua aplicabilidade pode ser realizada em diversos tipos de segmentos, como, por exemplo, mapear relacionamentos para descobrir associações em documentos digitais de processos penais de um determinado município.

Considera-se, assim, que é possível acompanhar a evolução de uma rede de relacionamentos, baseada em documentos, utilizando-se do período de criação dos documentos como parâmetro, mas o acesso as informações dos entes é fundamental para definir as aplicabilidade desses mapeamentos dentro do contexto social em que os arquivos que mapearam o grafo foram extraídos.

Notas

(1) Do inglês "rastreador web", um *web crawler* é um programa de computador que, de forma automatizada, acessa páginas da web a fim de coletar dados ou arquivos.
(2) O Apache Tika é uma ferramenta, escrita em Java, utilizada para extração de dados e texto de mais de mil tipos diferentes de arquivos, como os doc, docx e pdf encontrados no experimento. https://tika.apache.org
(3) Este conjunto de dados pode ser consultado pelo repositório: https://github.com/peluz/lener-br
(4) No contexto de Redes Complexas, quando nós possuem um grande número de conexões com outros nós, estes nós populares são denominados de *hubs* (Barabási & Bonabeau, 2003).

Referências

Agneessens, F., Borgatti, S. P.,and Everett, M. G. "Geodesic based centrality: Unifying the local and the global". *Social Networks*, n. 49, May 2017, pp. 12–26.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. "LeNER-Br: a dataset for named entity recognition in Brazilian legal text". In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), Canela, RS, Brazil. Springer. 2018, pp. 313–323,

Barabási, A. L., and Bonabeau, E. "Scale-Free Networks". *Scientific American*, n. 288, 2003, pp. 60–69. doi:10.1038/scientificamerican0503-60

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. U. "Complex networks: Structure and dynamics". *Physics Reports*, n. 424, 2006, pp. 175-308. doi:https://doi.org/10.1016/j.physrep.2005.10.009

Bollobás, B., Janson, S., and Riordan, O. "Sparse random graphs with clustering". *Random Structures & Algorithms*, n. 38, 2011, pp. 269–323.

Bonacich, P. "Power and Centrality: A Family of Measures". *American Journal of Sociology*, n. 92, 1987, pp. 1170–1182. doi:10.1086/228631

Borgatti, S. P. "Identifying sets of key players in a social network". *Computational & Mathematical Organization Theory*, n. 12, 2006, pp. 21–34.

Borgatti, S. P., and Everett, M. G. "A graph-theoretic perspective on centrality". *Social networks*, n. 28, 2006, pp. 466–484.

BRASIL. Constituição da República Federativa do Brasil. *Senado Federal*, 1988.

Clauset, A., Shalizi, C. R., and Newman, M. E. "Power-law distributions in empirical data". *SIAM Review*, n. 51, 2009, pp. 661–703. doi:10.1137/070710111

Coppin, B. *Inteligência Artificial*. Rio de Janeiro: LTC, 2017.

Fischetti, M. Physics *or Fashion? What Science Lovers Link to Most: Science aficionados have odd and surprising interests*. Scientific American, 2011, https://www.scientificamerican.com/article/graphic-science-science-lovers-web-traffic/

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. "Neural Architectures for Named Entity Recognition". *arXiv:1603.01360* [cs], 2016, http://arxiv.org/abs/1603.01360

Marteleto, R. M. "Redes Sociais, Mediação e Apropriação De Informações: situando campos, objetos e conceitos na pesquisa em Ciência da Informação". *Revista Telfract*, n. 1, 2018.

Mendonça, J., Macedo, H., Bisbo, T., Santos, F., Silva, N., and Barbosa, L. "Paramopama: a Brazilian-Portuguese corpus for named entity recognition". *12th National Meeting on Artificial and Computational Intelligence* (ENIAC) 2015.

Nadeau, D., and Sekine, S. "A survey of named entity recognition and classification". *Linguisticae Investigationes*, n. 30, 2007, pp. 3–26. http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002

Newman, M. E. "A measure of betweenness centrality based on random walks". *Social Networks*, n. 27, 2005, pp. 39-54. doi: https://doi.org/10.1016/j.socnet.2004.11.009

Pastor-Satorras, R., and Vespignani, A. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2007.

Presidência da República. *Detalhamento dos Servidores Públicos por Órgão, Portal da Transparência*, 2021, https://www.portaltransparencia.gov.br/servidores/orgao?ordenarPor=orgaoSuperiorLotacaoSIAPE&direcao=asc

Recuero, R. *Redes sociais na Internet*. Porto Alegre: Sulina, 2011.

Universidade Federal do Tocantins. Resolução nº 21, de 26 de outubro de 2016. *Guia de Redação e Formatação de Comunicações Oficiais*, 2016, pp. 18.

Wasserman, S., Faust, K., & others. *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press, 1994.

Watts, D. J. *Small worlds: the dynamics of networks between order and randomness*. Princeton university press, 2004.

Received: 30/09/2021 Accepted: 10/12/2022