Análise de redes de palavras baseada em títulos extraídos de um sistema de atendimento

Jansen Souza[1], Daniel Lyra[1], Julianne Cavalcanti, Rivaldo Simão, Zenildo Cézar, , Alexandre N. Duarte[1], Alisson V. Brito[1]

[1]Programa de Pós-Graduação em Informática - Centro de Informática – Universidade Federal da Paraíba (UFPB) – João Pessoa – PB – Brasil

{jansen.souza, dlyra4, rsym95, juliannecavalcanti, zenildo}@gmail.com, {alisson, alexandre}@ci.ufpb.br

Abstract. *Social networks Analysis (SNA) has been a topic of interest for many researches in last years. These networks can be established from the relation between people or the relation of information that can be analyzed in order to assist decision-making. The purpose of this article is to describe and analyze semantic networks constructed from the keywords present in trouble tickets titles submitted by users of a customer service system (Help Desk).Applying SNA techniques to an instance of this semantic network allowing the identification of the most common problems and facilitating the trouble shooting process.*

Resumo. *A análise de redes sociais (ARS) tem sido um tópico de interesse de diversos estudos realizados nos últimos anos. Essas redes podem ser estabelecidas a partir da relação entre pessoas ou da relação de informações que podem ser analisadas a fim de auxiliar a tomada de decisão. A proposta do presente artigo é descrever e analisar redes semânticas construídas a partir da relação de palavras-chave presente em títulos de chamados gerados por usuários de um sistema de atendimento (Help Desk). Aplicando as técnicas de ARS a uma rede semântica foi possível identificar os tipos de problemas mais comuns, favorecendo a tomada de decisão diante destes.*

1. Introdução

A principal motivação para a adoção de uma solução de análise de redes sociais que foque em interesses compartilhados é a busca por uma alta conectividade de seus componentes com a finalidade de promover um aumento na oportunidade de colaboração e compartilhamento de recurso, informações e conhecimentos [Fadigas et al. 2009]. Qualquer texto escrito pode ser transformado em uma rede de palavras, [Cunha et al. 2013]. Podemos modelar uma rede social, que tem como relacionamento entre seus atores, o vocabulário comum utilizado para compor determinados tipos de textos. As redes semânticas baseadas em títulos podem ser interpretadas como redes léxicas onde as palavras são nós e as arestas são as associações semânticas existente entre essas palavras [Pereira et al 2011]. Dentro dessa associação, formam-se um conjunto de vértices e arestas que conectam pares de palavras de um mesmo texto. Para analisar de forma mais específica essas conexões e relacionamentos entre vértices surge a Análise de Redes Sociais (ARS), que por sua vez, é amplamente adotada nos mais diversos contextos.

O objetivo geral é analisar o relacionamento entre as palavras utilizadas para descrever os problemas reportados em um sistema de *Help Desk,* utilizando *o* sistema em operação na Universidade XXX, como estudo de caso. Usuários de diversos setores geram solicitações com o objetivo de obter suporte e resolução para determinados problemas técnicos na área de Tecnologia da Informação. Cada solicitação possui um título que específica de forma breve o assunto do problema relatado pelo usuário. Esses títulos são compostos por dois campos, o setor de origem e a causa da solicitação. No presente trabalho, a partir da análise dos agrupamentos semânticos do texto em análise, buscamos identificar quais são as palavras mais recorrentes e a partir desse dado, traduzir em números a importância destas dentro do problema em questão. A rede de palavras construída nestes casos pode ser interpretada como um sistema de representação do conhecimento baseado em grafos. Os resultados apresentados permitem visualizar quais os problemas mais recorrentes a fim de facilitar a tomada de decisão e a definição de tendências nas solicitações de determinados usuários e/ou setores.

2. Redes Semânticas baseada em títulos (Estudo de Caso)

Neste experimento, exploramos os conteúdos dos títulos encontrados na base de dados do OTRS (Sistema de Chamados - *Help Desk* do Núcleo de Tecnologia da Informação – NTI da XXX). Nenhum desses títulos possui um tratamento prévio quanto à padronização de termos e setores de origem. Por este motivo, foi realizada uma mineração dos dados, para que a rede possa ser construída de forma organizada e padronizada. Todo esse processo é executado a partir de consultas na base de dados do OTRS alinhado ao uso de programas computacionais para a formação das redes.

O contexto desta pesquisa foca a análise dos relacionamentos entre palavras, para isso, procuramos respaldo na Análise de Redes Sociais para estudar a relação entre os termos mais mencionados. Contudo, utilizamos um método de construção de redes semânticas que consiste basicamente na eliminação das palavras sem significados intrínsecos (e.g. artigos, pronomes pessoais e possessivos, adjetivos possessivos, demonstrativos, interrogativos, advérbios, etc.) e na alteração das palavras restantes para sua forma padronizada. Para tanto, dois procedimentos principais foram necessários: o tratamento manual e o tratamento com o uso de programas computacionais. Na primeira parte, caracterizada como fase inicial da formação de dados, realizamos uma mineração a fim de detectar termos que não farão parte da rede. Baseado em [Fadigas et al. 2009] seguimos as seguintes regras de normalização:

a) foram removidas dos títulos, para uma formação clara do grafo, preposições, pronomes, advérbios, artigos, acentos e caracteres especiais. As conexões entre os vértices ocorrerão de acordo com a predominância das ligações existentes entre palavras-chave nos títulos dos chamados;

b) palavras repetidas no mesmo título foram excluídas, restando apenas uma ocorrência;

c) sequências de palavras com sentido único, devem formar uma única palavra, por exemplo, a expressão “portal da capes” será convertida em “portalcapes”. A palavra ou sentença é vista como a menor unidade de significado de um texto e cada palavra pode ter um sentido diferente a depender das palavras que estejam ao seu redor.

O tratamento com o uso de programas computacionais foi dividido em duas etapas. A primeira, focada diretamente no processo de identificação de padrões semânticos, para tratar questões como a eliminação das ambiguidades, palavras compostas, caracteres especiais, entre outros, mapeadas através do programa UNITEX, disponibilizado (open source) pela Rede Relex Brasil, que faz parte de um programa francês do Laboratoire d'Automatique Documentaire et Linguistique (LADL) [Caldeira 2006]. Por fim, através da identificação das palavras no UNITEX, foram executadas as regras mencionadas anteriormente, através de *scripts* (*shell script*), bem como geradas expressões regulares específicas, para substituição e/ou eliminação de determinadas palavras, objetivando assim uma padronização das palavras-chave.

Uma das padronizações realizadas foi referente aos setores ou unidades administrativas, por exemplo, "Centro de Informática", passou a ser "ci", com todas as palavras minúsculas.O exemplo a seguir, mostra um título de chamado sem tratamento:

"[Doutorado em Química CCEN] Email institucional para acesso ao Sci-finder"

Depois de realizada todas as etapas supracitadas, temos o seguinte título:

"ccen email institucional acessar scifinder"

Portanto, podemos verificar os atores ou vértices da rede, as palavras selecionadas por seus significados intrínsecos e os relacionamentos ou arestas que ocorrem entre as palavras de um título. As conexões entre os vértices ocorrerão de acordo com a predominância das ligações existentes entre palavras-chave nos títulos dos chamados.

Com os devidos ajustes no arquivo texto, foi criada uma aplicação em *Java* para gerar um novo arquivo no formato *Graph Modeling Language* (*GML)*, de modo que o programa utilizado em nossa pesquisa para análise (*Gephi*), monte a rede de palavras. É importante destacar que a rede aqui apresentada, foi elaborada considerando uma base de dados referente a um período de dois anos de uso na Instituição em estudo, contando, ao final, com um montante de 14.406 linhas de títulos dos chamados após realizarmos os filtros necessários com o objetivo de capturar as solicitações mais relevantes.

Destarte, foi gerada a rede de palavras-chave que ocorrem nos títulos, ligados entre si. Podemos verificar que algumas palavras-chave possuem uma maior incidência, tanto pela quantidade de arestas, quanto pelo tamanho do nó. As palavras "**email**", "**problema**" e "**acessar**" ocorrem com maior frequência, sendo assim, podem ser consideradas pontos centrais de interesse relacionados à problemas diante das solicitações dos usuários. A Figura 1 apresenta um dos componentes da rede formada.

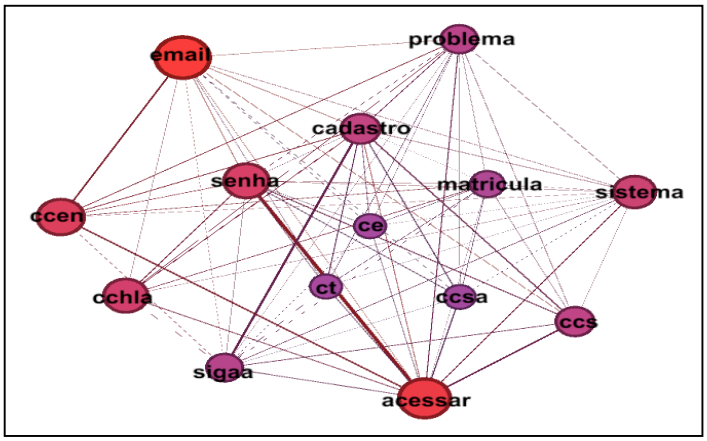

Figura 1. Rede formada após o processamento do arquivo *gml*

Destacamos três tipos de centralidade, grau (*degree*), proximidade (*closeness*) e de intermediação (*betweenness*). A centralidade de grau trata a importância de um vértice nas conexões que estabelece com vértices vizinhos e é quantificada pelo grau do vértice. Dessa forma, uma palavra apresenta um maior nível de importância, se estabelece um maior número de conexões com outros vértices vizinhos. Já a centralidade de proximidade evidencia o quanto um vértice está próximo de todos os demais vértices da rede. Este índice mostra a importância de uma palavra em relação aos vizinhos mais próximos, mas também a sua importância em relação a toda a rede de palavras, pois quanto mais central for um vértice numa rede, menor será sua distância para o vértice mais isolado da rede. E por fim, a centralidade de intermediação quantifica o número de vezes que um nó age como ponte ao longo do caminho mais curto entre dois outros nós.

3. Resultados

A rede gerada foi analisada e os resultados obtidos a partir do programa *Gephi*. A seguir apresentamos os principais resultados gerados a partir desta pesquisa. Os vértices representam cada palavra-chave e o peso de cada aresta, ligando dois vértices, representa quantas vezes duas palavras-chave apareceram juntas nos títulos dos chamados. Porém, é importante destacar que a simples quantificação não estabelece o mais importante nas análises de redes sociais que é a relação estabelecida entre as palavras. Alguns outros conceitos são utilizados, principalmente aqueles que denotam a importância das palavras em relação à centralidade.

A Tabela 1 mostra todos os quantitativos relativos ao número de solicitações por setor ou unidades administrativas, e seus respectivos índices de centralidade. Conforme já exemplificado anteriormente, cada chamado possui dois campos, um relativo ao setor que originou a solicitação e outro relativo ao problema que motivou a origem do chamado. Consequentemente, de posse desses resultados, a direção do NTI poderá facilmente identificar que, por exemplo, o CCEN (Centro de Ciências Exatas e da Natureza) foi o centro que mais gerou ordem de serviço, ou seja, problemas referentes à área de T.I., em um período de dois anos. Outra análise interessante que podemos destacar é que boa parte dos problemas relacionados a este último centro, estão diretamente ligados com a palavra-chave email.

Tabela 1.: Quantitativos por setor e seus respectivos índices de centralidade.

	Índices						
Setor	**Número de Vértices**	**Número de arestas**	**Número de títulos**	**Número de palavras**	**Degree**	**Closeness**	**Betweenness**
CCEN	902	1895	1489	5708	351	2,239	463.176
CCHLA	891	1867	1273	4963	327	2,262	410.055
CCS	782	1739	1289	5176	278	2,325	298.971
CE	598	1147	726	2742	224	2,398	199.061
CI	273	459	236	940	105	2,535	42.450
CT	629	1359	902	3579	224	2,373	186.894

A Tabela 2 mostra as ligações, não direcionadas, entre palavras dos chamados e o peso entre elas. O peso denota a quantidade de vezes que essa ligação existiu, podendo ser chamado também de densidade da aresta. Podemos evidenciar, a partir desse estudo, que a ligação origem-destino, por exemplo, entre as palavras-chave email e institucional, apresenta o maior número de ocorrência, destacando assim que, boa parte dos problemas referentes à T.I. estão diretamente ligados ao contexto "**email institucional**" e com isso, ao concentrar esforços sobre este, diminuir especificamente cada ponto de estrangulamento de uma determinado setor.

Tabela 2.: Relação dos pesos em relação às arestas que conectam os vértices.

Origem/Destino	**Origem/Destino**	**Peso**
Email	institucional	426
Portalcapes	acessar	423
Acessar	Senha	331

De acordo com a Figura 2, seguindo o critério de centralidade de grau, as palavras "email" e "acessar" apresentam uma maior frequência de ligações, evidenciando sua ocorrência nos títulos dos chamados. Através dos resultados gerados pelo *Gephi*, pode-se concluir que as palavras "**email**", "**senha**" e "**acessar**" são aquelas mais importantes, do ponto de vista dos três tipos de centralidades abordados neste trabalho. A interpretação da análise de redes sociais aplicada às redes de títulos mostra que as três palavras acima são aquelas que estão mais próximas de suas vizinhas na rede, bem como mais próximas de todas as outras (Fadigas et al. 2009).

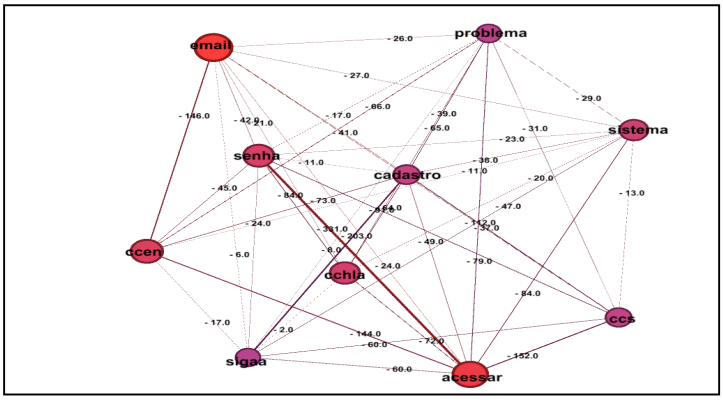

Figura 2. Rede formada com as informações de pesos cada aresta

4. Considerações Finais

O presente trabalho apresenta uma relevante contribuição para empresas que necessitem, por exemplo, de um sistema de gerenciamento de serviços para centralizar, auxiliar a tomada de decisão e resolver as solicitações mais recorrentes de seus respectivos clientes. O uso da metodologia de análise de redes sociais juntamente com as características de exploração visual disponibilizadas na ferramenta *Gephi*, permitiram e facilitaram a combinação de elementos para entendimento e avaliação do modo de comunicação, troca de informações e conhecimentos na comunidade estudada. O padrão da comunicação dos vértices, ou palavras-chave, responsáveis pela dinamização da rede, foi devidamente avaliado e identificado.

A visão fornecida pela ferramenta *Gephi* complementou as informações geradas através da base de dados do sistema de *Help Desk,* utilizado no Núcleo de Tecnologia da Informação da XXX, pois introduziu uma visão das palavras-chave mais mencionadas nos títulos dos chamados por setor das solicitações de atendimento ao usuário. A pesquisa mostrou que o núcleo de problemas mais recorrentes nos títulos dos chamados, de acordo com o grau de centralidade, compõe-se das palavras-chave "**email**", "**acessar**" e "**senha**", na ordem de maior incidência para menor respectivamente, bem como, os contextos, "**email institucional**" e "**portalcapes acessar**", possuem os maiores pesos em suas arestas.

Contudo, através deste mapeamento, podemos destacar que o uso de redes semânticas baseadas em títulos, no geral, auxilia no entendimento da integração das palavras-chave. Desta forma, pode-se apontar evidências de como a alta gestão poderá atuar no processo de tomada de decisão sobre possíveis soluções ou tendências, no que diz respeito aos problemas que ocorrem com maior frequência em um determinado local.

Referências

[Cunha et al. 2013] Cunha, M. V. ; Rosa, M. G. ; Fadigas, Inácio de Sousa ; Miranda, J. G. V. ; Pereira, H. B. B. . Redes de títulos de artigos científicos variáveis no tempo. In: XXXIII Congresso da Sociedade Brasileira de Computação, 2013, Maceió. CIDADES INTELIGENTES: DESAFIOS PARA A COMPUTAÇÃO. Rio de Janeiro: SBC, 2013. v. 1. p. 1744-1755.

[Caldeira 2006] Caldeira, S. M. G.; Petit Lobão, T. C.; Andrade, R. F. S.; Neme, A. e Miranda, J. G. V. (2006). The network of concepts in written texts. The European Physical Journal B, v. 49, pp. 523-529.

[Fadigas et al. 2009]Fadigas, I., Henrique, T., Pereira, H., Senna, V., and Moret, M. (2009). Analise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. Educação Matemática, Pesquisa, 11(1):167–193.

[Pereira et al. 2011]Pereira, H., Fadigas, I., Senna, V., and Moret, M. (2011). Semantic networks based on titles of scientific papers. Physica A: Statistical Mechanics and its Applications, 390(6):1192 – 1197.