Similaridade entre Documentos de Especificação de Requisitos de Software Utilizando o Modelo Vetorial de Recuperação de Informação

Adriana Cássia da COSTA (1) Custódio Gastão da SILVA JUNIOR (2)
(1) IFMT, Rua Professora Zulmira Canavarros, adriana.org@gmail.com
(2) IFMT ,Rua Professora Zulmira Canavarros, gastaojunior@gmail.com

RESUMO
O Processamento de Linguagem Natural ganhou destaque nos últimos anos, principalmente por propor soluções as limitações existentes na manipulação de grandes volumes de informações textuais. Este trabalho aborda o estudo do modelo vetorial de recuperação de informação na busca por termos relevantes que consigam indicar a similaridade entre documentos, com o foco na utilização desse método pela engenharia de requisitos. Foi realizado um estudo de caso para analisar o comportamento do método TFIDF (*Term frequency – Inverse Document Frequency*) em documentos de especificação de requisitos de software. Os testes demonstraram que na aplicação do método TFIDF foram obtidos resultados com considerável similaridade conceitual entre os documentos. Os documentos utilizados nos testes fazem parte de um domínio de um sistema acadêmico e os resultados mostram que o método TFIDF juntamente com a seleção lingüística de substantivos e verbos torna o processo de busca por similaridade conceitual entre documentos mais eficiente.

Palavras Chave: TFIDF, software, requisitos, documentos, informação

1. Introdução
O Processamento de Linguagem Natural – PLN está entre os grandes desafios da computação, pois envolve a junção dos conhecimentos linguísticos com os computacionais para prover mecanismos “inteligentes” e eficientes de interação homem-máquina. Algumas contribuições importantes estão na área de tradução automática que evoluiu consideravelmente nos últimos anos na recuperação de informação e na engenharia de software, nesta última, com estudos de aplicações visando principalmente a etapa de especificação de requisitos de software.
Em engenharia de software, alguns trabalhos como de [Sayão (2007)] e [Carvalho et al (2007)], estudam a aplicação de técnicas de PLN na verificação e validação de requisitos com o objetivo de melhorar a qualidade e auxiliar o trabalho das equipes na análise de requisitos. A maioria das especificações de requisitos é escrita em linguagem natural, em alguns casos sendo complementadas por outros tipos de notações como diagramas, equações, modelos formais, etc [Silva e Martins]. Atualmente é comum que o desenvolvimento de software seja feito por equipes geograficamente distribuídas, para estes casos, a verificação e validação de requisitos são de alta complexidade, podendo envolver o tratamento de até milhares de requisitos [Sayão(2007)]
As técnicas de PLN associadas ao levantamento de requisitos permitem a detecção de erros durante a fase inicial de especificação do projeto possibilitando a economia de tempo e trabalho. [Sayão(2007)] afirma que técnicas de PLN podem ser aplicadas a documentos de requisitos e outros artefatos, apoiando no processo de validação e verificação de requisitos. A similaridade entre documentos de especificação de requisitos permite que sejam extraídos conceitos importantes para permitir, por exemplo, a verificação de regras de negócios duplicadas e o a manutenção de padrões de especificação de requisitos.
Este trabalho demonstra a utilização do método vetorial de recuperação de informação pelo método TFIDF (*Term frequency – Inverse Document Frequency*) aplicados a documentos de especificação de requisitos de software. O estudo de caso visa à busca por similaridade conceitual entre estes documentos. O artigo está organizado do seguinte modo: Na Seção 2, são descritos os conceitos básicos sobre PLN, na Seção 3 é apresentado um resumo sobre engenharia de software e o papel do documento de especificação de requisitos nessa área,

na Seção 4 é apresentado o método utilizado para a recuperação de informação e na Seção 5, é apresentado o estudo de caso desenvolvido e as considerações finais.

2. Conceitos e Aplicações do Processamento de Língua Natural

Entre os benefícios do Processamento de Língua Natural é possível mencionar os avanços na Biomedicina, como a estruturação de ontologias como a Foundational Model of Anatomy que segundo [Schulz e Freitas (2009)], fornece conhecimento declaratório sobre a estrutura microscópica do ser humano. Dentre outras utilidades do PLN está o acesso a bases de dados, recuperação de informações precisas, além de sistemas de suporte a decisão. O Processamento de Língua Natural apresenta estudos tanto para a língua escrita como na língua falada, os estudos são feitos principalmente nas áreas de tradução, interpretação, geração e tradução de textos, requisições e consultas, reconhecimento e síntese da fala, entre outros citados por [Nunes (2009)].

Uma importante área do PLN é a recuperação de informação, que visa principalmente trabalhar com informações relevantes para o usuário no contexto em que ele está inserido. Nas etapas para a extração de informação relevante é encontrada a dificuldade no processamento semântico das informações, já que o computador não possui a capacidade de inferir conclusões que seriam simples para os seres humanos. Estes possuem a percepção para diferenciar significados de termos como “dente”, que tanto pode ser dente de alho, dente de um ser humano ou até mesmo de alguma ferramenta de oficina mecânica, dependendo do contexto em que for inserido. Por este motivo, é necessário criar mecanismos que possibilitem alcançar melhores resultados, enriquecendo semanticamente os mecanismos computacionais como, por exemplo, estabelecendo os critérios de seleção lingüística.

A seleção lingüística é a etapa onde são selecionadas as classes gramaticais relevantes para o domínio em questão que serão utilizadas como base para a aplicação do método. Neste trabalho, foram considerados apenas os substantivos e verbos para a aplicação do método TFIDF, as demais classes gramaticais foram desconsideradas, formando a *stoplist* (lista de termos ignorados no processamento). Na busca por termos relevantes que sejam capazes de representar o conhecimento de um domínio existem 3 abordagens normalmente utilizadas, a lingüística que considera as informações sintáticas, semânticas e morfológicas; a estatística que normalmente considera a freqüência de ocorrência de termos e a híbrida que trata da junção das duas técnicas anteriores e que atualmente tem apresentado resultados significativos na busca por termos relevantes como no caso do tratamento de informações para a extração de informação com a finalidade de preparar os dados para a sumarização automática [Pereira et al (2002)].

É um consenso na área de processamento de linguagem natural que os métodos de extração de termos podem ser agrupados segundo a abordagem utilizada em: linguísticos e estatísticos. No entanto, esta divisão raramente é estanque, pois praticamente todos os métodos tem ao menos algum componente de cada uma das abordagens. Métodos baseados em informações linguísticas sempre levam em consideração algum critério de frequência, assim como métodos estatísticos usualmente consideram algumas listas de palavras que seguem critérios linguísticos. [Lopes et al (2009)]. Após essa apresentação dos principais pontos para a compreensão da atuação do PLN, é importante ressaltar que de acordo com [Conteratto (2006)], a eficiência dos sistemas de PLN está diretamente relacionada com a qualidade das informações linguísticas.

3. Engenharia de Software

Na década de 80 surgiram as metodologias de desenvolvimento de software com a finalidade de suprir a necessidade de organizar e estruturar de maneira eficiente o desenvolvimento e possibilitar o atendimento a grandes demandas. Uma característica comum aos paradigmas do desenvolvimento de software conceituados por [Sommerville(2003)] e [Pressman(1995)] é a fase de obtenção de requisitos para a produção da documentação do software, dessa etapa inicial resulta o primeiro documento do projeto.

O documento de especificação de requisito possui as informações necessárias para orientar os diversos usuários envolvidos no projeto, como engenheiro de software, cliente e desenvolvedor. Os requisitos, de acordo com [Larman (2000)], descrevem as necessidades ou desejos para um produto com o objetivo de identificar e documentar as informações. Esse documento não possui um padrão em especifico, podendo variar em sua forma de acordo com as necessidades do projeto, no entanto, deve esclarecer as restrições e

atribuições do sistema. A vantagem da linguagem natural é a facilidade de comunicação entre os atores envolvidos, pois não necessita de treinamento específico para sua compreensão. [Silva e Martins] A contribuição de [Silva e Martins], está relacionada ao aperfeiçoamento da engenharia de requisitos, com um estudo sobre a ferramenta PARADIGMA. Esta ferramenta pretende auxiliar o engenheiro de requisitos através da geração do modelo conceitual de classes a partir dos requisitos em linguagem natural que ainda precisa de ajustes feitos pelo engenheiro. De acordo com os autores o diferencial da ferramenta está no uso de padrões linguísticos que permitem aproximar os modelos de classes gerados de maneira automatizada dos modelos criados por modeladores humanos.
A contribuição de [Tardelli (2005)], trata do estudo do modelo vetorial TFIDF com o uso da técnica de Trigram Phrase Matching como alternativa a normalização. Os resultados obtidos no trabalho demonstram que o método pode ser aplicado para atribuir conceitos médicos a textos da área da Saúde, nos experimentos deste autor são atribuídos aos documentos similares das fontes de informação do estudo de caso do LILACS/MEDLINE.
Este trabalho é semelhante aos estudos de [Silva e Martins], [Catarina(2009)] e [Lopes et al (2009)], no que diz respeito a melhoria dos processos da engenharia de software. No entanto, os estudos de [Silva e Martins], [Catarina (2009)] e [Lopes et al (2009)], tem como base a análise de desempenho de ferramentas visando a automatização dos processos. Este estudo tem como foco a análise do funcionamento do método TFIDF (*Term frequency – Inverse Document Frequency*), aplicado a documentos de especificação de requisitos de software.

4. Recuperação de Informação

A recuperação de informação visa o processamento de informações relevantes para o usuário, sendo utilizada principalmente em grandes volumes de dados. A aplicação de conceitos lingüísticos em mecanismos de busca começou na década de 70, antes disso, eram bastante difundidos modelos como o booleano, que utiliza expressões lógicas nas consultas, exigindo assim certo nível de conhecimento do usuário. Os modelos linguísticos são considerados mais simples por não utilizarem “termos técnicos” para melhoria dos resultados ou necessitar de treinamento do usuário.
Existem diversos modelos para a recuperação de informação, no entanto, o foco deste trabalho está no modelo vetorial que trabalha com o critério de relevância para a classificação dos termos do domínio. [Gonzalez (2003)], afirmam que o modelo vetorial tem grande influência no desenvolvimento de sistemas de recuperação de informação, este modelo utiliza mecanismos que aperfeiçoam resultados da pesquisa através da aplicação de conceitos de linguística na manipulação dos dados.

4.1 Método TFIDF (*Term frequency – Inverse Document Frequency)*

Este método surgiu na década de 80 e foi desenvolvido por Salton e Buckey, trata-se da junção de duas medidas de freqüência de termos, a medida TF (Term Frequency) determina a freqüência de ocorrência simples do termo no documento, essa medida também é chamada de peso local, pois resultará no valor do peso para os termos em um único documento. Considerando a medida TF, se em um documento o termo “escolher” ocorrer 5 vezes, então é atribuído o peso 5 a este termo.
A medida IDF (*Inverse Document Frequency)* verifica a freqüência de ocorrência do termo em relação a todos os documentos do domínio, com essa medida reduz os casos de falsos positivos[1], quando termos de pouca relevância são apontados como representativos do domínio. A medida IDF contribui para a obtenção de melhores resultados já que trabalha com a freqüência de ocorrência do termo no conjunto de documentos do domínio, dessa maneira os termos apontados como relevantes não são necessariamente os que têm maior freqüência de ocorrência. Neste caso, se em um conjunto composto por 10 documentos e analisando o termo “escolher”, é avaliada a ocorrência no conjunto de 10 documentos antes de ser atribuído o peso do termo. [Sayão (2007)] afirma que essa junção minimiza os pesos dos termos que aparecem com muita freqüência nos documentos, pois estes geralmente não são considerados representativos de significado para um domínio, normalmente nessa categoria estão: às conjunções, artigos, pronomes. Os resultados após a aplicação do

[1] Pode acontecer em casos onde artigos como “a”, “o” aparecem com freqüência maior, serem considerados mais relevantes que termos representativos para o contexto, como os verbos e substantivos, isso ocorre quando estes tiverem menor frequência que o primeiro, pois a comparação entre as frequências vai apontar como relevante os artigos, ou seja, um falso positivo .

método TFIDF podem atingir valores infinitos e para facilitar a comparação dos resultados é indicador normalizá-los para uma forma que facilite a análise dos resultados.

$$tfidf(t_k, d_j) = tf(t_k, d_j) \cdot \log \frac{|\mathcal{D}|}{\#_{\mathcal{D}}(t_k)}$$

Figura 1 Medida TFIDF

De acordo com [Lavelli et al (2009)], o método TFIDF é representado pelo produto entre a medida TF e a IDF como mostrado na figura 2. Nesta fórmula tk representa a freqüência de ocorrência de um termo específico no documento e dj a freqüência de ocorrência do termo na quantidade total de documentos do domínio. Na segunda parte da fórmula |D| representa a quantidade total de documentos do domínio e #D(tk) é a quantidade total de documentos onde o termo é encontrado. Além disso, é acrescentada a função log a essa medida, [Sayão (2007)] justifica que neste caso, o uso da função log evita que termos que aparecem em somente um documento tenham pesos maiores que termos que aparecem em dois documentos do conjunto.
A aplicação desse método prevê ainda a aplicação de uma medida para normalizar os resultados em uma forma em que os resultados sejam valores restringidos, facilitando a análise dos resultados, neste trabalho foi utilizada a medida do cosseno da similaridade.

$$sim = \frac{\sum_{i=1}^{n} (x_{ki} \cdot y_{ki})}{\sqrt{\sum_{i=1}^{n} (x_{ki})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_{ki})^2}}$$

Figura 2 Medida do Cosseno da similaridade

A medida do cosseno da similaridade faz com que os resultados das comparações entre dois documentos sejam normalizados e seus valores fiquem entre 0 e 1. Os pares de documentos com taxa de similaridade próxima a 1 são muito similares e quando as taxas são próximas a 0 os documentos são pouco similares.
Na Figura 3 demonstra como é realizada a comparação entre os termos de dois documentos, na primeira parte o W*ki* representa os pesos dos termos do vetor de do primeiro documento chamado de x e Y*ki* são os pesos dos termos de outro documento chamado de y. Nesse método os documentos são considerados “vetores de termos”, ou seja, cada documento é um vetor que pode ter uma quantidade indeterminada de termos. A comparação acontece sempre entre pares de documentos verificando os termos em comum, ou seja, são analisados dois vetores de termos por vez e os pontos em comum são os termos compartilhados por ambos os vetores.

5.0 Modelo Vetorial de Recuperação

No modelo vetorial é feita a atribuição de pesos aos termos de busca, esses pesos são usados como medida de relevância entre os termos do domínio. Os documentos são representados em uma matriz de termos onde os documentos e cada termo contido neles é colocado em linhas e colunas como representado na Figura a seguir:

	ser	**devem**	**podem**	**escolher**
Manter_Calendario	6	3	2	6
Manter_Cursos	2	5	5	3
Manter_Áreas	2	1	3	3
Manter_Req_Inscrição	1	2	4	1

Tabela 1 Matriz de representação do modelo vetorial

Na Figura 1, é representado um exemplo do modelo de matriz de termos para um domínio, a primeira coluna corresponde aos documentos do domínio, da segunda coluna em diante, são dispostos todos os termos que ocorrem nestes documentos. No exemplo acima, foram considerados os termos com frequência de ocorrência a partir de um, pois palavras menos frequentes podem ter grande relevância para documentos de especificação de requisitos.

O valor 6 na posição (2,2) representa o peso do termo “ser” no documento Manter_Calendario, significando que neste documento a palavra “ser” ocorre 6 vezes. O peso dos termos é atribuído pela frequência de ocorrência dele no documento, neste caso o termo “ser” tem peso 6, pois é a quantidade de vezes este termo aparece no documento Manter_Calendário.
Este modelo considera o cálculo de similaridade para facilitar a comparação entre os documentos, neste trabalho foi utilizada a medida do cosseno da similaridade que restringe os valores entre 0 e 1. De acordo com este método, os documentos com valores de similaridade próximos a 1 possuem alta similaridade e se o resultado se aproximar de 0, os documentos apresentam baixa similaridade.

5.1 Estudo de Caso

Foi realizado um experimento para testar o modelo vetorial aplicando o método TFIDF que será processado para indicar o nível de similaridade conceitual entre os documentos. A similaridade conceitual possivelmente será encontrada em documentos que compartilham o mesmo domínio. Em documentos de um mesmo domínio é esperado que os termos considerados relevantes consigam representar o conteúdo existente nesses documentos.
A seleção lingüística para este estudo foi composta pelos verbos e substantivos do fluxo básico e das regras de negócio desses documentos, por considerar que esses itens possuem uma quantidade considerável de termos relevantes para o contexto e assim melhorar o desempenho do método TFIDF. O domínio da aplicação contou com 28 documentos de especificação de requisitos de software separados em dois conjuntos de teste, o primeiro com 6 documentos com verbos e 6 documentos com substantivos do fluxo básico, no segundo conjunto de teste foram selecionados 8 documentos de verbos e 8 documentos de substantivos das regras de negócio

Figura 4: Visão Geral das Etapas do Processamento

Após a divisão dos dois conjuntos, foi necessário criar a *stoplist,* uma lista com os termos que serão desconsiderados durante o processamento, os termos dessa lista recebem o nome de *stopwords*. Se as *stopwords* fossem processadas poderiam influenciar no resultado de modo à tendenciar a resultados inconsistentes, pois ocorrem com alta freqüência e não são relevantes para o domínio em questão.
Para a seleção por freqüência simples de ocorrência dos verbos e substantivos, foi utilizada a ferramenta contador de freqüência simples Lácio Web desenvolvida pelo Núcleo Interinstitucional de Linguística Aplicada – NILC[2]. Algumas observações devem ser colocadas sobre este trabalho, foram considerados os termos com freqüência de no mínimo uma ocorrência por documento para tentar alcançar o maior número de termos relevantes. Os resultados dos experimentos conseguiram identificar documentos similares, para o conjunto dos verbos do fluxo básico, as taxas de similaridade acima de 0,2 expressavam documentos com maior similaridade no domínio como na Figura abaixo, uma comparação entre dois documentos com similaridade 0, 277.

Manter_Credenciamento_Professores_E	**Manter_Calendário_F**
Fluxo Básico Usuário Solicita Manter Credenciamento de Professores Este caso de uso inicia quando o usuário da PRPPG solicita a opção “Manter “Credenciamento de Professores”. Então, o usuário tem as seguintes opções:	**Fluxo Básico** Usuário Solicita Manter Calendário Este caso de uso inicia quando o usuário da PRPPG, Secretaria de Programa ou Coordenadoria de Programa solicita a opção “Manter Calendário”. Então, o usuário tem as seguintes opções:

[2] NILC – Núcleo Interinstitucional de Linguística Aplicada (http://www.nilc.icmc.usp.br/nilc/)

Cadastrar um Credenciamento (dispara o **Sub-fluxo de Inclusão**). - Pesquisar os dados de um Credenciamento (dispara o **Sub-fluxo de Consulta**) O usuário seleciona uma opção. Usuário seleciona a opção desejada de cadastrar, alterar, consultar ou excluir Credenciamento. O sistema executa o sub-fluxo correspondente.	- Cadastrar uma Data (dispara o **Subfluxo de Inclusão**). Configurar uma Fase (dispara o **caso de uso "Instanciar Fase e Requisitos de Andamento"**). - Definir parâmetros de calendário (dispara o **Sub-fluxo de Alteração de Parâmetros de Calendário)**. - Pesquisar os dados de uma Data (dispara o **Sub-fluxo de Pesquisa**) O usuário seleciona uma opção. Usuário seleciona a opção desejada de cadastrar, alterar, consultar ou excluir Data. O sistema executa o sub-fluxo correspondente.

Tabela 2 Comparação entre documentos com taxa de similaridade 0, 277

Nesse primeiro conjunto de teste, foi possível concluir que as taxas de similaridade com valore de 0, 277 e acima deste valore foram melhores para expressar os documentos similares, nas taxas abaixo desse valor os documentos continham pouca similaridade como no na tabela 2 que mostra a comparação entre documentos com taxa de similaridade 0, 175.

Consultar_Situação_Inscritos _	**Manter_Credenciamento_Professores_E**
Fluxo Básico **Pré-condições** · Usuário está autenticado e possui acesso a esta funcionalidade. **Fluxo Básico** Usuário solicita Consultar Situação de Inscritos O sistema mostra a tela de filtros a serem preenchidos. O usuário informa os filtros e solicita a opção Pesquisar. O sistema lista os inscritos que obedecem os filtros escolhidos. Então, o usuário pode alterar os dados de um inscrito (dispara o **Sub-fluxo de Alteração**)	**Fluxo Básico** Usuário Solicita Manter Credenciamento de Professores Este caso de uso inicia quando o usuário da PRPPG solicita a opção "Manter Credenciamento de Professores". Então, o usuário tem as seguintes opções: Cadastrar um Credenciamento (dispara o **Subfluxo de Inclusão**). - Pesquisar os dados de um Credenciamento (dispara o **Sub-fluxo de Consulta**) O usuário seleciona uma opção. Usuário seleciona a opção desejada de cadastrar, alterar, consultar ou excluir Credenciamento. O sistema executa o sub-fluxo correspondente.

Tabela 3 Comparação entre documentos com taxa de similaridade 0.175

Para o conjunto de testes dos substantivos do fluxo básico, a maioria dos documentos apresentou taxas de similaridade abaixo de 0,1. Neste caso como a classe gramatical selecionada foi dos substantivos e isso tornarem a quantidade de termos maior em relação ao conjunto de testes dos verbos, pode ter influenciado para que houvesse similaridade entre os documentos até mesmo para as taxas menores. A tabela 3 demonstra esse tipo de caso que apresenta um documento desse conjunto com taxa de similaridade de 0, 091.

Manter_Cursos_D	**Manter_Credenciamento_Professores_E**
Fluxo Básico Usuário Solicita Manter Cursos Este caso de uso inicia quando o usuário da PRPPG solicita a opção "Manter Cursos". Então, o usuário tem as seguintes opções: - Pesquisar os dados de um Curso (dispara o **Sub-fluxo de Consulta**) O usuário seleciona uma opção. Usuário seleciona a opção desejada de alterar ou consultar Curso. O sistema executa o sub-fluxo correspondente.	**Fluxo Básico** Usuário Solicita Manter Credenciamento de Professores Este caso de uso inicia quando o usuário da PRPPG solicita a opção "Manter Credenciamento de Professores". Então, o usuário tem as seguintes opções: - Cadastrar um Credenciamento (dispara o **Sub-fluxo de Inclusão**). - Pesquisar os dados de um Credenciamento (dispara o **Sub-fluxo de Consulta**) O usuário seleciona uma opção. Usuário seleciona a opção desejada de cadastrar, alterar, consultar ou excluir

	Credenciamento. O sistema executa o sub-fluxo correspondente.

Tabela 4 Comparação entre documentos com taxa de similaridade 0, 091

No conjunto de teste II a aplicação do método TDIDF foi realizada somente nos verbos e substantivos das regras de negócio. As taxas de similaridade com boa similaridade foram os valores acima de 0, 048 para os verbos e 0,188 para os substantivos. Abaixo está um documento desse conjunto de teste com taxa de similaridade na faixa de valor que apresentou bons resultados.

Manter_areas_B	**Manter_linhas_pesquisa_F**
[RN01] Os campos área e ativa são obrigatórios. **[RN02]** Valor de domínio que indica os status das áreas (Suspensa, Ativa, Cancelada). Aqui,"Suspensa" indica que a área irá ser cancelada mas ainda tem alunos em curso e Cancelada é a área que não possui mais alunos cursando. **[RN03]** Apresentar "Ativa" como opção default. **[RN04]** Lista de Cursos ativos no sistema de Pós-Graduação e que pertencem ao mesmo Programa da Área sendo editada. **[RN05]** Não podem existir duas áreas com um mesmo nome para o mesmo Programa. **[RN06]** Devem ser apresentados somente os Programas que estão ativos no sistema. **[RN07]** Quando for removida a área de concentração, então não pode mais ser apresentada e nem armazenada no sistema a informação de disponibilização daquela área para todos os cursos.	**[RN01]** Valor de domínio que indica as linhas de pesquisa que estão ativas ou inativas (Sim, Não). **[RN02]** Apresentar "Sim" como opção default. **[RN03]** Devem ser apresentados todos os cursos que estiverem ativos no sistema, no entanto as Coordenadorias de Programa visualizam apenas os cursos do seu programa associado. **[RN04]** As opções apresentadas em "Área" devem ser filtradas pela opção escolhida no campo "Programa". **[RN05]** Não podem existir duas linhas de pesquisa com um mesmo nome para a mesma Área de Concentração. **[RN06]** Lista de todos os Programas disponíveis no sistema de Pós-Graduação Stricto Sensu, no entanto as Coordenadorias de Programa visualizam apenas o seu programa associado. **[RN07]** Quando for removida a linha de pesquisa ou a sua associação com a área, então não pode mais ser apresentada e nem armazenada no sistema a informação de disponibilização daquela linha para a área a qual estava relacionada.

Tabela 5 Comparação entre documentos com taxa de similaridade 0, 326

Na Tabela 5 está um exemplo de documento apresentado boa similaridade, nesse conjunto de teste os resultados encontrados abaixo de 0, 326 apresentavam pouca similaridade conceitual. Os documentos de cada conjunto foram comparados ao pares entre si, a taxa de similaridade representa sempre o resultado da comparação entre os termos comum em pares de documentos.

6. Considerações Finais

Os resultados obtidos com este trabalho sugerem que a utilização do método TFIDF (Term frequency – Inverse Document Frequency) é eficiente na procura por termos relevantes para a representação conceitual dos documentos. A escolha gramatical dos substantivos permitiu melhores resultados para os documentos do fluxo básico, no entanto, para as regras de negócio, os documentos analisados a partir dos verbos apresentaram melhores resultados no sentido de facilitar a análise dos resultados e encontrar documentos que realmente apresentavam similaridade.

Acredita-se que as manipulações de informações lingüísticas em conjunto com o desenvolvimento de software possam contribuir para a melhoria na qualidade e rendimento das equipes de desenvolvimento. Na área de especificação de requisitos de software, a aplicação das técnicas de PLN é promissora e o objetivo é utilizar os resultados de estudos como este em projetos maiores para a melhoria do processo de desenvolvimento de software. Em alguns estudos como de [Sayão (2007)], contam com medidas estatísticas para automatizar parcialmente o processo de verificação de duplicidade de requisitos. Deste modo, é possível acreditar que estudos como este possam colaborar para a melhoria dessas técnicas e conseqüentemente da etapa de especificação de requisitos de software.

A descoberta da similaridade entre documentos de especificação de requisitos de software pode auxiliar o processo de correção de requisitos, principalmente para os casos como os relatados por [Sayão (2007)], onde os desenvolvedores trabalham separadamente na elaboração dos documentos.

7. Referências

CARVALHO, G. SAYÃO, M. GATTI, M. **Técnicas de PLN na Análise de Domínio em SMAs Abertos**. PUC-Rio.2007.

CONTERATTO, G. B. H. **Semântica e Computação: uma interação necessária para o aperfeiçoamento de sistemas PLN**. Letras de Hoje. Porto Alegre. V.41, nº 02, p.353-367, junho 2006

DIAS-DA-SILVA, B. C. et al. **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações**. Agosto, 2007.

GONZALEZ, M. LiIMA, V. L. S. **Recuperação de Informação e Processamento da Linguagem Natural**. XXIII Congresso da Sociedade Brasileira de Computação, Campinas, 2003. Anais do III Jornada de Mini-Cursos de Inteligência Artificial, Volume III, p.347-395.

LARMAN, C. **Utilizando UML e Padrões: uma introdução à análise e ao projeto orientado a objetos**; Trad: Luiz A. Meirelles Salgado. Porto Alegre: Bookman. 2000.

LAVELLI, A. SEBASTIANI, F. ZANOLLI,R. **Distributional Terms Representations: An Experimental Comparison**. 2004.

NUNES, G. **Desafio do Processamento de Linguas Naturais** – Workshop Projeto Farol. Porto Alegre, Março 2009.

PRESSMAN, R. S. **Engenharia de Software.** Markon Books: 1995.

RICH, E. KNIGHT, K. **Inteligência Artificial**. 2.ed. São Paulo: Makron Books,1993.

SAYÃO, M. **Verificação e Validação em Requisitos: Processamento da Linguagem Natural e Agentes**. Rio de Janeiro, Abril 2007. PUC-Rio.

SCHULZ, S. FREITAS, F. **RECIIS Revista Eletrônica em Informação & Inovação em Saúde**. Rio de Janeiro, v.3, n.1, Março 2009.

SOMMERVILLE, I. **Engenharia de Software**. Tradução: André Maurício de Andrade Ribeiro. São Paulo: Addison Wesley. 2003.

SILVA, W. C. da, MARTINS, L. E. G. **PARADIGMA: Uma Ferramenta de Apoio à Elicitação e Modelagem de Requisitos Baseada em Processamento de Linguagem Natural**. 11th Workshop on Requirements Engeneering.

LOPES, L, OLIVEIRA, L. H. M., VIEIRA, R. **Análise Comparativa de Métodos de Extração de Termos: Abordagens Linguística e Estatística**. Porto Alegre. 2009

CATARINA, R, S. **Extração de Casos de uso no Processo de Análise de Software através de Análise Sintática e Semântica em um Documento de Requisitos**. PUCRS, Porto Alegre. 2009.

TARDELLIi, A. **Implementação do Método Trigram Phrase Matching para problemas de similaridade de textos**. 4ª Reunião de Coordenação Regional da BVS grupo de trabalho TI. Salvador. 2005.