MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2016/08.05.18.28-RPQ

ESTUDOS DE MECANISMOS DE INTEROPERABILIDADE ENTRE BANCOS DE DADOS COM TECNOLOGIAS DISTINTAS - FASE 2: PESQUISA E APLICAÇÃO DE MEDIDAS DE SIMILARIDADE

Lise Christine Banon

Programa de Capacitação Institucional – PCI/ MCT/ INPE.
Relatório Final de Atividades

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34P/3M7SBQ8>

INPE
São José dos Campos
2016

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2016/08.05.18.28-RPQ

ESTUDOS DE MECANISMOS DE INTEROPERABILIDADE ENTRE BANCOS DE DADOS COM TECNOLOGIAS DISTINTAS - FASE 2: PESQUISA E APLICAÇÃO DE MEDIDAS DE SIMILARIDADE

Lise Christine Banon

Programa de Capacitação Institucional – PCI/ MCT/ INPE.
Relatório Final de Atividades

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34P/3M7SBQ8>

INPE
São José dos Campos
2016

PROGRAMA DE CAPACITAÇÃO INSTITUCIONAL – PCI/ MCT/ INPE

RELATÓRIO FINAL DE ATIVIDADES

Nome do bolsista: Lise Christine Banon

Orientador da bolsa: José Carlos Neves Epiphanio

Período de vigência da bolsa: 01/10/2008 a 30/09/2010

Modalidade da bolsa: DTI/7E

RELATÓRIO FINAL DE ATIVIDADES

Título do projeto científico:

Estudos de mecanismos de interoperabilidade entre bancos de dados com tecnologias distintas - Fase 2: Pesquisa e aplicação de medidas de similaridade.

1) Histórico

Atualmente as grandes instituições de ensino e pesquisa como, por exemplo, Universidade de São Paulo (USP), Universidade Estadual de Campinas (UNICAMP), Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) e Instituto Nacional de Pesquisas Espaciais (INPE) possuem sistemas próprios de armazenamento de dados bibliográficos.

Há um grande interesse em permitir a comunicação entre esses bancos. Um exemplo disso é o IBICT que, por meio do protocolo de coleta de metadados "Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)" colocou à disposição na sua Biblioteca Digital de Teses e Dissertações (BDTD) as referências de teses e dissertações destas grandes instituições (TRISKA e CAFÉ 2001).

Em relação as referências bibliográficas, o Brasil dispõe de um banco nacional conhecido como Plataforma Lattes, do CNPq, que foi desenvolvido pelo grupo STELA (PPGEP/ UFSC – Programa de Pós-Graduação em Engenharia de Produção - Universidade Federal de Santa Catarina), e é um resultado da concepção conjunta do MCT, CNPq, CAPES e FINEP, para integração dos sistemas de informação em Ciência e Tecnologia no Brasil.

Desde 1995, o INPE dispõe de uma Biblioteca Digital que hospeda toda a sua produção técnico-científica (BANON 2006). Esta biblioteca digital, organizada segundo o paradigma de repositórios uniformes, denominada Uniform Repository for a Library (UR*Lib*), conta com um banco de referências bibliográficas programada na linguagem TCL (Tool Command Language).

A primeira fase deste projeto, ocorrida no período de 2006 a 2008, viabilizou o desenvolvimento de um sistema de importação de referências da Plataforma Lattes para a biblioteca digital do INPE, um grande avanço no processo de indexação de referências, trazendo benefícios diretos à Instituição e à comunidade científica.

No entanto, as análises dos resultados das primeiras importações indicaram que o potencial da ferramenta desenvolvida poderia ser prejudicado, devido ao considerável número de duplicidades identificadas nesta primeira fase do projeto.

Por um lado os bibliotecários contaram com a garantia da rápida coleta de boa parte da produção científica da Instituição; e por outro, a desvantagem de um imenso trabalho manual para identificar e excluir todas as redundâncias. A correção das inconsistências foi

cuidadosa e precisou ser elaborada em um curto prazo, pois a demora poderia comprometer a precisão dos indicadores da produção científica institucional, que são entregues semestralmente ao MCT.

Na Plataforma Lattes, não há um identificador persistente para todas as referências, o que dificulta o controle das referências importadas, possibilitando a ocorrência de duplicidades no acervo Institucional.

Na UR*Lib*, cada referência bibliográfica recebe um identificador persistente e uma chave de citação, que consiste em uma cadeia de caracteres criada a partir de trechos dos metadados da referência (como sobrenome do autor, ano e título). Esta *string* foi criada apenas para ser usada na ordenação dos resultados de busca da UR*Lib*. No entanto, logo no início do projeto, alguns estudos indicaram que esta chave de citação também poderia ser utilizada nos processos de importação, minimizando as ocorrências de duplicidades no banco. De fato, o seu uso impossibilitou casos pontuais de duplicidade durante as importações, bem como evitou os casos de entradas subseqüentes de uma referência com erros de cadastro, que porventura já tivesse sido importada e normalizada no acervo institucional. Apesar destes benefícios, a chave de citação não foi capaz de evitar um considerável número de duplicações.

Portanto, a segunda fase deste projeto foi dedicada à pesquisa de medidas de similaridade para tentar identificar métricas que possam solucionar o problema de redundância entre dados bibliográficos. A conclusão desta etapa do projeto resultou na proposta do uso de um aplicativo para a rápida detecção e exclusão dos casos de duplicações de dados bibliográficos no acervo institucional, a partir de uma medida de similaridade.

2) Resumo do projeto

As etapas descritas inicialmente no Plano de Trabalho Detalhado sofreram algumas alterações. Houve modificações, no que diz respeito à seqüência, inclusão e substituição de algumas etapas.

Etapas descritas inicialmente no Plano de Trabalho Detalhado:

- Levantamento das possíveis medidas de similaridades;
- Estudo do Sistema de Importação de dados da Plataforma Lattes para a UR*Lib*;
- Montagem de um grupo controlado de dados bibliográficos;
- Estudo estatístico do comportamento das regras de decisões usando cada uma das medidas selecionadas;
- Inclusão da melhor medida de similaridade no Sistema de Importação existente;
- Publicação de um artigo e redação do relatório final.

Etapas desempenhadas efetivamente neste Projeto:

- Montagem de um grupo controlado de dados bibliográficos;
- Levantamento das possíveis medidas de similaridades;
- Estudo da Plataforma UR*Lib* e do Sistema de Importação de dados da Plataforma Lattes para UR*Lib*;
- Desenvolvimento de um Sistema de Recomendação;
- Estudo estatístico do comportamento das regras de decisões usando a medida de similaridade selecionada;
- Proposta de uma interface web para gerenciar as referências similares na Biblioteca Digital: Similar Search UR*Lib* Accessory;
- Redação do relatório final.

Considerações sobre as alterações:

- A seqüência de execução das três primeiras etapas foi modificada. Encontram-se na literatura várias medidas de similaridade e cada uma apropriada para uma determinada aplicação. Portanto, iniciar a pesquisa pela classificação dos dados e pelo estudo da natureza dos erros que levavam as duplicações foi fundamental para selecionar uma medida de similaridade que atendesse aos objetivos deste Projeto.
- A inclusão da etapa "Desenvolvimento de um Sistema de Recomendação" foi utilizada para a avaliação do comportamento do grupo controlado de dados bibliográficos e a definição do grau crítico de similaridade.
- O item originalmente "Estudo do Sistema de Importação de dados da Plataforma Lattes para a UR*Lib*" foi substituído por "Estudo da Plataforma UR*Lib* e do Sistema de Importação de dados da Plataforma Lattes para UR*Lib*". Esta alteração é mais apropriada, pois funções específicas foram estudadas em ambos, para viabilizar o desenvolvimento das etapas subseqüentes, sem comprometer o desempenho dos processos já existentes.
- A utilização da medida de similaridade no Sistema de Importação existente evitaria os principais casos de duplicações, mas também poderia prejudicar as importações de referências semelhantes (não duplicadas), com alto grau de similaridade. Desta forma, ao substituir esta etapa, optou-se por uma opção mais segura, introduzindo uma "Proposta de interface web para gerenciar as referências similares na Biblioteca Digital: Similar Search UR*Lib* Accessory".
- O artigo a ser submetido, no qual serão apresentadas as metodologias desenvolvidas neste projeto, está em sua fase inicial.

3) Objetivo

O INPE possui um sistema de indexação bibliográfica denominado UR*Lib*, que gerencia a avaliação, busca e análise da produção bibliográfica do Instituto. O desenvolvimento do sistema de importação de dados bibliográficos da Plataforma Lattes para a Plataforma UR*Lib* enfrenta uma série de problemas, que exigem grande retrabalho de conferência para cada referência bibliográfica devido às limitações da Plataforma Lattes para o processo de importação.

A finalidade desta pesquisa – que se configura como um avanço em relação à anterior - é estudar soluções para o problema de identificação dos dados bibliográficos que não possuem identificadores persistentes. Mais precisamente, o objetivo é estudar medidas de similaridades entre dados bibliográficos, e conseqüentemente, regras de decisão (NEVEU, 1970) para reconhecer se dois dados bibliográficos são ou não referentes à mesma obra.

4) Atividades desenvolvidas durante o período da bolsa

4.1) Montagem de um grupo controlado de dados bibliográficos

No final da primeira fase deste projeto, ainda no ano de 2008, a Biblioteca Digital do INPE iniciou o processo de importação de referências da Plataforma Lattes. Na segunda fase do projeto, foram analisados os resultados das importações ocorridas nos anos de 2008 e 2009.

Os resultados obtidos neste período podem ser observados na **Tabela 1**:

Tabela 1 - Comparativo entre o total de aquisições na Biblioteca Digital (nas categorias artigos em revista, trabalhos em evento, capítulos de livro e livros) em relação ao total de importações (aprovadas e reprovadas), nos anos de 2008 e 2009.

REFERÊNCIAS (REF)	ANO 2008			ANO 2009		
	Nº	% REF/TI	% REF/TA	Nº	% REF/TI	% REF/TA
Total de Aquisições (TA)	1524			1539		
Total de Importações (TI)	615			668		
Aprovadas (IA)	533	87%	35%	536	80%	35%
Reprovadas (IR)	82	13%		132	20%	

O Serviço de Informação e Documentação (SID) - INPE oferece vários recursos para a recuperação da produção científica institucional. Além da importação de referências da Plataforma Lattes, há outros sistemas como: Auto-arquivamento, Arquivamento pelo SID, Publicação Digital de Anais do INPE, Sistema de Coleta de Dados das Secretarias e Sistema de E-print.

Para uma comparação efetiva entre o número de referências importadas e o número de aquisições obtidas por todos os sistemas citados acima, o item TA da tabela refere-se ao total de aquisições nas mesmas categorias de referências consideradas em TI. Conforme orientação do SID, apenas quatro categorias de referências são importadas da Plataforma Lattes: artigos em revista, trabalhos em evento, capítulos de livro e livros. As demais categorias de documentos, como por exemplo: as teses e dissertações possuem um processo específico de publicação, no qual os profissionais da informação acompanham os diversos estágios do documento, orientando os autores em cada etapa e utilizando recursos da Biblioteca Digital como histórico de atualização, controle de versões e restrição de acesso externo, sendo desnecessária a sua importação.

Outro detalhe importante em relação à tabela: todos os dados apresentados referem-se estritamente às publicações com pelo menos uma autoria INPE. Portanto, as fontes utilizadas para expor os valores que constam em TA foram: os Indicadores da Produção Científica do INPE ano 2008 (versão publicada em 27/02/2009 - 13h21) e ano 2009 (versão publicada em 12/03/2010 - 10h29). No caso das importações, apenas as publicações de autoria INPE são mantidas no acervo institucional. Sendo assim, os valores do total de importações (TI), bem como os seus subitens: importações aprovadas (IA) e importações reprovadas (IR) foram obtidos consultando a Biblioteca Digital.

Se as importações sem autoria INPE foram excluídas do acervo, este foi um dos motivos para parte das referências importadas serem reprovadas. O sistema de importação de referências da Plataforma Lattes adotado pelo INPE, tem permissão para utilizar um serviço restrito do CNPq que possibilita o acesso aos currículos dos pesquisadores no formato XML. Se o pesquisador responsável pelo currículo tiver cadastrado o INPE como afiliação, quando o sistema fornecer o CPF deste pesquisador, o acesso ao arquivo XML deste currículo será autorizado. Desta forma, em 2008, além dos servidores públicos, o SID solicitou aos bolsistas, discentes, estagiários e celetistas que cadastrassem o INPE na afiliação de seus currículos. Esta colaboração trouxe bons resultados, pois em dois anos consecutivos, o sistema de importação foi responsável pela recuperação de 35% do total de aquisições (TA). As publicações vinculadas às teses, dissertações e aos projetos de pesquisa, em sua maioria ocorrem após a Defesa e a conclusão da Bolsa de Pesquisa. Sendo assim, o SID solicitou que em 2009 os CPF dos discentes e bolsistas fossem mantidos no sistema de importação, mesmo nos casos em que o candidato houvesse concluído o curso de pós-graduação ou a bolsa de pesquisa. Isso facilitou a recuperação de várias referências vinculadas ao INPE, mas também importou referências externas. Em 2008, o sistema de importação continha inicialmente: 2.086 CPF, e na última importação em 2009, este número subiu para 2838.

De fato, na tabela é possível observar que em 2008 a percentagem de IR por TI correspondia a 13% e este valor subiu para 20% no ano de 2009.

Além das referências sem autoria INPE, o motivo que levou o SID a descartar parte das referências importadas foram os casos de duplicações.

Logo abaixo, a **Tabela 2** exemplifica esta afirmação, ao apresentar os valores das importações reprovadas no ano de 2009, na categoria artigo em revista.

Tabela 2 - Importações de artigos em revista reprovadas no ano de 2009 e os percentuais ao subdividi-las em duas categorias: sem autoria INPE ou duplicas, conforme o motivo da reprovação.

REFERÊNCIAS (REF)	ANO 2009 Artigo em Revista	
	Nº	% REF/IR
Importações Reprovadas (IR)	43	
⇒ Sem autoria INPE	14	33%
⇒ Duplicadas	29	67%

As referências reprovadas pelo SID compunham um rico material de estudo para a segunda fase deste projeto. Utilizando um recurso da Biblioteca Digital, as referências importadas e selecionadas como reprovadas pelos especialistas do SID, não foram excluídas do sistema, mas ocultas das interfaces acessadas pelo usuário final**.**

O estudo destas referências ocultas proporcionou a montagem do grupo controlado de dados bibliográficos. Após classificar as importações reprovadas em duas categorias: sem autoria INPE e duplicadas; em cada par de referências duplicadas registrou-se o identificador persistente da UR*Lib* de ambos. Nos casos em que uma referência oculta estava triplicada ou quadruplicada, ainda assim, os similares foram analisados aos pares. O ID permitiu extrair os metadados das referências e comparar os valores dos campos em cada par de similares. Se os valores dos campos de um par de similares divergiam, a ocorrência era especificada em uma planilha.

Ao selecionar uma pequena amostra de uma das planilhas (os 25 registros iniciais da planilha contendo a avaliação das referências na categoria artigo em evento no ano de 2009) e restringindo-se apenas na análise dos campos AUTORIA e TÍTULO, é possível exemplificar algumas inconsistências observadas e o número de suas ocorrências, conforme apresentado na **Tabela 3**:

Tabela 3 - Referências de artigos em eventos reprovadas no ano de 2009 e os percentuais das inconsistências observadas nos campos de metadados: Autoria e Título.

	ANO 2009 Artigos em Eventos		
	REFERÊNCIAS (REF)	Nº	% REF/IRD
	Importações Reprovadas Duplicadas Analisadas (IRD)	25	
AUTORIA	Formatação ou acentuação incorreta	5	20%
	Erro na grafia de pelo menos um autor	4	16%
	Ausência de pelo menos um autor	7	28%
	Inversão na ordem de autoria	8	32%
	Inversão na ordem dos nomes de pelo menos um autor	5	20%
TITULO	Formatação ou acentuação incorreta	7	28%
	Erro na grafia de uma das palavras do título	3	12%
	Ausência de alguma palavra do título	1	4%

Nesta tabela, cada linha dos campos autoria e título definem um tipo de inconsistência, ao qual é associado o número de inconsistências encontradas e sua percentagem. Nas 25 referências analisadas, verificou-se que 28% possuem simultaneamente pelo menos mais de um tipo de inconsistência.

A primeira inconsistência descrita como: formatação ou acentuação incorreta (nos campos autoria e título) causou duplicações no início das importações. Logo na primeira fase deste projeto adotou-se um recurso existente na UR*Lib*, a inclusão de chaves de citação para todas as referências importadas, com o intuito de minimizar as duplicações, A chave de citação é uma cadeia de caracteres criada a partir de trechos dos metadados da referência (como sobrenome do autor, ano e título) e foi criada apenas com o intuito de facilitar a ordenação das referências no resultado da busca. A introdução deste recurso no sistema de importação trouxe benefícios, mas logo no início da segunda fase deste projeto, foi necessário ajustar o módulo do sistema responsável pela criação desta chave de citação eliminando a sensibilidade às letras maiúscula e à acentuação, para que não houvessem casos de duplicação devido a este tipo de inconsistência.

De acordo com as análises realizadas para os dados de 2008 e 2009, as inconsistências ocorreram com menos freqüência no campo TÍTULO. Por exemplo, ao desconsiderar os casos de formatação ou acentuação incorreta (que atualmente não causam duplicação) da pequena amostra acima, tem-se que 24% das referências restantes apresentam alguma inconsistência no campo título, enquanto 94% apresentam alguma inconsistência no campo autoria, sendo que 18% apresentam inconsistências em ambos os campos.

Durante as análises, as buscas pelo duplicado da referência oculta eram efetuadas manualmente a partir do sistema de busca da Biblioteca Digital. Ao alternar diferentes palavras do título nas expressões de busca, observou-se a recuperação de grande parte das referências duplicadas. Esta estratégia motivou o desenvolvimento de um Sistema de Recomendação, associando buscas avançadas da Biblioteca Digital ao cálculo de similaridade. Os sistemas de recomendação sugerem ao usuário um conteúdo semelhante ao encontrado a partir de um sistema de busca, diminuindo o seu esforço para localizar informações de interesse, por isso atualmente estes sistemas são muito usais na internet. O sistema de recomendação foi desenvolvido neste projeto, com o intuito de avaliar o comportamento do grupo controlado de dados bibliográficos, permitindo a escolha do grau crítico de similaridade, i.e., o valor que corresponde ao limiar entre uma referência ser considerada o duplicado ou apenas semelhante à referência alvo (terminologia adotada neste texto, para identificar uma referência que a partir do sistema de recomendação, recebe uma lista de referências selecionadas pelo sistema de busca como similares a esta).

4.2) Levantamento das possíveis medidas de similaridade.

Duas possíveis medidas de similaridades foram consideradas.

Considerou-se a primeira medida de similaridade com base na correlação de *Pearson* (PAPOULIS, 1990) e a segunda com base na maior – subseqüência – comum (CORMEN et al., 2001).

A primeira medida de similaridade se apresentou como uma proposta mais abrangente, pois o seu domínio de aplicação é mais amplo. A segunda medida de similaridade apresentou sensibilidade à ordem das palavras, enquanto a primeira podia ou não ser sensível a ordem das palavras, dependendo de como fossem construídas ou escolhidas as duas variáveis aleatórias.

Neste trabalho, foi escolhida a primeira abordagem, por ser mais flexível e por não ser necessariamente sensível a ordem.

4.3) Estudo da Plataforma UR*Lib* e do Sistema de Importação de dados da Plataforma Lattes para UR*Lib*;

Funções específicas do Sistema de Importação de dados da Plataforma Lattes e UR*Lib* foram revistas e estudadas a fundo, para viabilizar o desenvolvimento das etapas subseqüentes, sem comprometer o desempenho dos processos já existentes.

Por exemplo, no caso do Sistema de Importação de dados da Plataforma Lattes, a proposta inicial consistia em incluir o processo do cálculo de similaridade durante as importações, com o intuito de evitar duplicações no acervo. Sendo assim, funções referentes ao uso da chave de citação (fundamental na primeira fase deste projeto para evitar casos específicos de duplicações) precisaram ser revistas para que a inclusão do novo processo no Sistema de Importação não alterasse a funcionalidade da chave de citação.

No caso da UR*Lib*, também podemos citar um exemplo: durante o planejamento do Sistema de Recomendação, constatou-se que o retorno da busca pelos similares não poderia adotar a interface padrão do retorno da busca da UR*Lib*, pois não teria sentido o Sistema de Recomendação apresentar as dez referências mais recentes, e sim as mais similares. Este pequeno detalhe, demandou pesquisa e estudo das funções pré-existentes na UR*Lib*, pois a nova interface precisava ser desenvolvida com base nos moldes da interface padrão, sem comprometer as diversas funções pré-existentes e interligadas a este processo.

4.4) Desenvolvimento de um Sistema de Recomendação

A **Figura 1** apresenta esquematicamente o diagrama de fluxo de dados do Sistema de Recomendação desenvolvido.

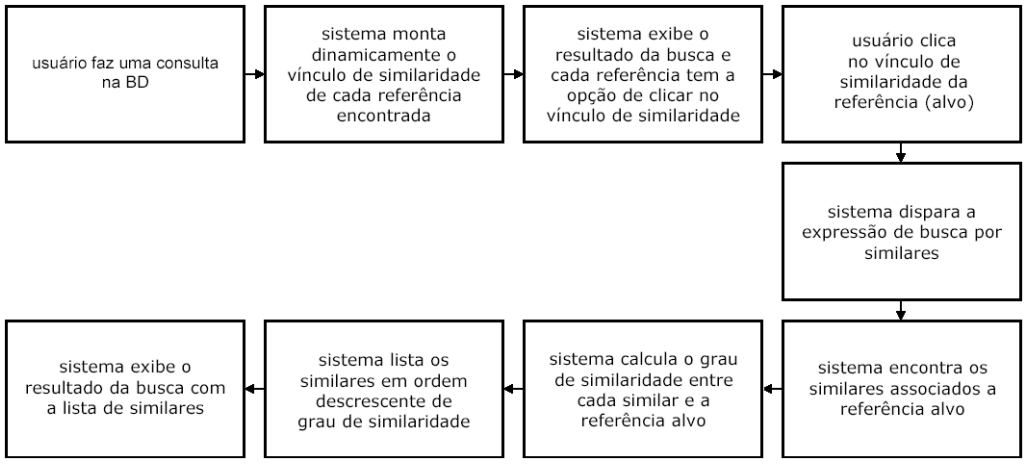

Figura 1 - Diagrama de Fluxo de Dados do Sistema de Recomendação

A **Figura 2,** logo abaixo, exemplifica o funcionamento do sistema de recomendação:

Figura 2 - Interfaces relacionadas ao funcionamento do Sistema de Recomendação.

Inicialmente o usuário consulta à Biblioteca Digital, digitando no campo de busca as suas palavras-chaves, que neste exemplo foram: formaggio 2009. Ao clicar em buscar (item 1 da **Figura 2**), o sistema dispara a expressão de busca e ao encontrar as referências que satisfazem esta expressão, gera dinamicamente para cada referência, a opção de acesso aos "similares", cujo link neste texto recebeu a terminologia de vínculo de similaridade. Assim que todos os vínculos de similaridade são gerados, o sistema exibe a interface com o resultado da busca.

No item 2 da **Figura 2**, o usuário acessa o conteúdo do vínculo de similaridade de uma das referências do resultado da busca, que neste exemplo representa a referência alvo. A seta em azul na **Figura 2** aponta para a interface do resultado da busca dos similares da referência alvo. No início desta interface, item 3 da **Figura 2**, está a expressão de busca do vínculo de similaridade. Logo abaixo, em cor de fundo diferenciada, consta a referência alvo, sempre exibida na primeira posição para permitir que, em uma única interface, o usuário possa facilmente comparar os dados da referência alvo em relação aos seus similares. Em seguida, as referências similares se apresentam ordenadas, em prioridade do maior grau de similaridade. Em cada referência listada nesta interface, consta o grau de similaridade em relação à referência alvo. Por exemplo, o item 4 da **Figura 2**, indica o grau de similaridade da própria referência alvo, cuja similaridade é máxima, ou seja, similaridade: 1.00.

4.4.1) Desenvolvimento do vínculo de similaridade

4.4.1.1) Parâmetros do vínculo de similaridade

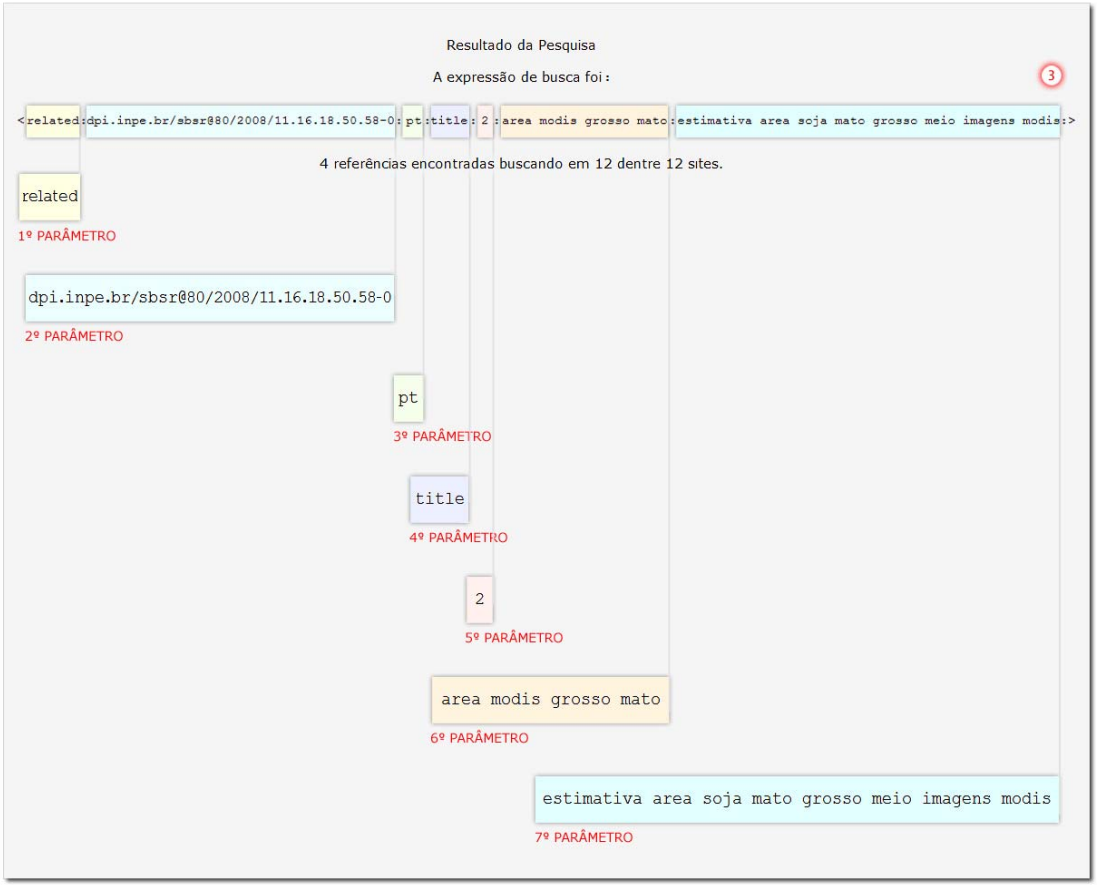

Figura 3 - Parâmetros da estrutura do vínculo de similaridade (item 3 da figura 2, em detalhe).

O vínculo de similaridade é formado pela concatenação de vários parâmetros. A seguir, constam as identificações de cada parâmetro, as justificativas de sua inclusão, bem como o processo ao qual estão relacionados: busca por referências, busca por similares, cálculo de similaridade e apresentação do resultado da busca. A identificação de cada parâmetro da estrutura do vínculo de similaridade é apresentada na **Figura 3** (detalhe da área especificada no item 3 da Figura 2).

⇒ **1º Parâmetro:** Related

A estrutura do vínculo de similaridade inicia-se por este parâmetro chamado: "related".

A existência deste parâmetro estabelece a exibição da interface do resultado da busca pelos similares, ao invés da interface do resultado da busca padrão. Além disso, sempre que um registro satisfizer uma expressão de busca iniciada por este parâmetro, o cálculo de similaridade é efetuado.

⇒ **2º Parâmetro:** Repositório de metadados da referência alvo

Este parâmetro interfere na apresentação da interface do resultado da busca pelos similares.

O repositório de metadados da referência alvo é o parâmetro que possibilita programar a interface do resultado da busca pelos similares, posicionando a referência alvo em primeiro lugar e diferenciando a sua cor de fundo em relação às demais referências. O intuito é facilitar a comparação entre o alvo e seus similares, sem a necessidade de acessar duas interfaces distintas.

Por serem dinâmicos, os vínculos de similaridade apresentam como vantagem: a atualização do conteúdo acessado em tempo real. No entanto, podem ocorrer casos em que o resultado da expressão de busca é nulo, ou seja, nenhuma referência é encontrada. Quando o resultado de um vínculo de similaridade é nulo, o ideal seria não exibi-lo na interface. Para este fim, o sistema precisaria testar o resultado de cada vínculo e retirá-lo nos casos em que não houvesse ocorrências. No entanto, este processo não deve prejudicar a velocidade de retorno do resultado da busca, ou seja, a exibição desta interface não pode aguardar o retorno do teste de todos os vínculos, pois isso comprometeria o desempenho do sistema. Uma alternativa futura para a reformulação do resultado da busca seria desenvolver um script no cliente que, após a exibição do resultado da busca, testasse os vínculos e os retirasse da interface quando não houvesse ocorrências. Nesta fase do projeto, não houve reformulação do resultado da busca. Atualmente, a partir deste parâmetro, qualquer expressão de busca disparada por um vínculo de similaridade retornará pelo menos uma ocorrência - a própria referência alvo.

Na interface de resultado da busca, nem todas as referências apresentam um vínculo de similaridade associado. A ausência pode ocorrer por dois motivos:

- sem preenchimento do campo idioma na referência ou

- título muito sucinto (quando o número de palavras do título, com potencial para ser pré-processado pelo sistema como relevante, for inferior a quatro).

⇒ **3º Parâmetro:** Idioma

O valor deste parâmetro influi na busca pelos similares. A partir do idioma cadastrado, são definidas as inflexões gramaticais em número das palavras importantes usadas na montagem da expressão de busca pelos similares.

Ao prever a inclusão das inflexões em número nas expressões de busca pelos similares, estes casos passaram a ser detectados pelo sistema.

Para as referências que não possuíam o campo idioma preenchido, os testes indicaram que seria inviável aplicar um programa para inflexão em número, pois cada idioma possui regras gramaticais específicas e diversas exceções.

As referências importadas da Plataforma Lattes sempre apresentam o preenchimento do campo idioma. Desta forma, foi possível desenvolver o sistema para que as inflexões em número fossem consideradas no cálculo da similaridade, ao menos para o idioma português e inglês.

⇒ **4º Parâmetro:** Campo(s) de metadado(s) cujos valores serão selecionados como relevantes na expressão de busca.

O valor deste parâmetro estabelece os campos de metadados dos quais o sistema deve extrair as palavras que serão consideradas como relevantes na expressão de busca pelos similares. O programa que define a estrutura do vínculo de similaridade pode ser personalizado, ou seja, o administrador pode definir estes campos. De acordo com a análise do grupo controlado de dados bibliográficos, o mais indicado foi definir apenas o campo título como parâmetro. Caso futuramente seja preciso detectar a similaridade em outras aplicações, extraindo informações relevantes em outros campos de metadados, apenas com a definição de um novo valor para este parâmetro, o programa estará personalizado para outra aplicação.

⇒ **5º Parâmetro:** Número de combinações das palavras relevantes

O valor deste parâmetro estabelece o número de combinações das palavras relevantes, na montagem da expressão de busca pelos similares.

De acordo com os testes, os melhores resultados foram obtidos a partir de expressões de busca que continham duas combinações com n-1 elementos cada, onde n é o número de palavras relevantes e, neste caso adotou-se n=4. A primeira combinação consiste em associar as três primeiras palavras mais importantes e a segunda combinação consiste em associar a primeira, a segunda e a quarta palavra mais importante.

Por exemplo, se quatro palavras a, b, c, d estão nesta ordem decrescente de relevância, as combinações seriam: abc e abd

⇒ **6º Parâmetro:** Palavras relevantes do(s) campo(s) definido(s) no 4º parâmetro

Neste parâmetro constam as palavras relevantes usadas na montagem da expressão de busca pelos similares. A seleção destas palavras é feita a partir do título, o campo de metadado definido no 4° parâmetro, como o mais indicado.

Definiu-se a relevância de cada palavra do título, a partir do cálculo de sua freqüência nos campos: título, palavras-chave e resumo.

Para utilizar as palavras do campo título, resumo e palavras-chave, foi necessário desenvolver um módulo para o pré-processamento dos dados, que basicamente possuía as seguintes finalidades:

- tratar inconsistências resultantes de erros de digitação (como por exemplo: a inclusão de chaves, colchetes, hífen e etc);

- eliminar a sensibilidade às letras maiúscula e à acentuação, ou seja, tornar-se "*case-insentive*" e "*accent-insensitive*"; e

- eliminar palavras sem significado expressivo, definindo uma lista de palavras que deveriam ser desconsideradas do cálculo de similaridade (como por exemplo, os artigos e pronomes).

Este recurso de seleção de palavras relevantes foi testado exaustivamente. Para cada grupo de dados testado, foi elaborado um relatório descrevendo as inconsistências observadas. A partir do relatório, as rotinas de pré-processamento sofriam os ajustes necessários, seguidos de novos testes para a confirmação do bom desempenho do sistema, nos itens apontados como críticos nos relatórios de análise. O módulo de pré-processamento teve impacto em outros processos, como por exemplo: a opção de tornar a seleção de palavras relevantes insensível à acentuação ("accent-insensitive") repercutiu no desenvolvimento das rotinas para as inflexões, pois algumas regras gramaticais baseiam-se na acentuação.

Para obter as palavras relevantes considerou-se a importância das palavras, diretamente proporcional a sua freqüência de ocorrência.

Nos casos em que a freqüência de ocorrência coincidiu, o critério de seleção do sistema optou pela palavra mais extensa.

⇒ **7º Parâmetro:** Palavras do título da referência alvo

Este último parâmetro armazena as palavras do título da referência alvo, com a finalidade de utilizá-las no cálculo da similaridade, em relação a cada registro encontrado a partir da expressão de busca do vínculo de similaridade, como uma referência similar ao alvo.

O vínculo de similaridade poderia ser considerado um link pré-programado, pois para a sua criação foi desenvolvido um algoritmo no qual é definida a estrutura lógica que gerencia previamente o seu conteúdo. Este algoritmo define a concatenação dos parâmetros descritos acima, em seguida os valores destes parâmetros são extraídos dinamicamente dos metadados da referência, e a partir da estrutura lógica que foi definida para a montagem da expressão de busca, obtém-se cada vínculo de similaridade.

4.4.1.2) Montagem da expressão de busca do vínculo de similaridade

A montagem da expressão de busca do vínculo de similaridade segue os seguintes passos, exemplificados logo abaixo:

a) Definição da inflexão de cada palavra relevante de acordo com o idioma:

language = pt

importantWordList = imagem soja area vegetacao

inflectionList = {imagem imagens} {soja sojas} {area areas} {vegetacao vegetacoes}

b) Montagem da busca truncada de cada par (no padrão do comando *glob* em *TCL*).

inflectionList = image[mn]* soja* area* vegetac[ao][oe]*

c) Montagem das duas primeiras combinações de n-1 palavras dentro de n palavras.

combinationList = { image[mn]* soja* area*} { image[mn]* area* vegetac[ao][oe]*}

d) Montagem da expressão de busca.

searchExpression = {title, image[mn]* and title, soja* and title, area*} or

{title, image[mn]* and title, soja* and title, vegetac[ao][oe]*}

Atualmente a expressão de busca por similares baseia-se na seleção de quatro palavras relevantes do título da referência alvo. A relevância é definida a partir do cálculo da freqüência de ocorrência das palavras do título, em outros campos como, por exemplo, resumo e palavras-chaves. A partir do idioma, são definidas as inflexões das palavras selecionadas como relevantes. Atualmente a expressão de busca é formada a partir de duas combinações de três das quatro palavras relevantes, considerando-se também as suas inflexões.

O conteúdo de um vínculo de similaridade não é armazenado no banco de dados, mas processado a cada retorno da busca, o que permite a atualização dos vínculos em tempo real. Isso significa que ao haver a correção de uma referência na biblioteca digital, alterando o valor de um campo utilizado para gerar a expressão de um vínculo de similaridade, assim que a atualização for efetuada, qualquer usuário que consultar esta referência na biblioteca digital, poderá verificar que o vínculo de similaridade associado a esta referência também estará atualizado. Da mesma forma, se houver a inserção de um grande número de referências na biblioteca digital (como por exemplo, pelas importações da Plataforma Lattes), automaticamente estas novas referências ficarão acessíveis a partir dos vínculos de similaridade relacionados a estas referências.

A questão do retorno da busca estar atrelado à conclusão da montagem dos vínculos de similaridade, não afetou o desempenho do sistema. A seleção de valores e cálculos efetuados para a montagem dinâmica dos vínculos de similaridade ocorre apenas em relação aos metadados da referência, portanto o processo não prejudicou a velocidade de retorno da busca.

Ainda nos casos em que a expressão de busca resultava em um elevado número de referências, como a Biblioteca Digital dispõe de um recurso para a exibição do resultado de busca por pacotes, constatou-se a rápida exibição da interface com o resultado da busca.

4.4.1.3) Propriedade da busca pelos similares

A montagem da expressão de busca depende dos valores preenchidos em diferentes campos de metadados. Portanto, se A for um similar de B, B não será necessariamente um similar de A. Sendo assim, pode-se afirmar que a busca pelos similares possui a propriedade não-comutativa.

4.4.2) Cálculo de similaridade

Ao clicar em um vínculo de similaridade, o sistema retorna a lista dos similares priorizando a exibição das referências que possuem maior grau de similaridade em relação à referência alvo. Para que isso aconteça, na fração de segundo em que o usuário clica no vínculo de similaridade até a exibição dos similares, o sistema calcula dinamicamente o grau de similaridade do alvo em relação a cada similar (encontrados a partir da expressão de busca do vínculo de similaridade), ordena e em seguida exibe na interface do cliente a lista dos similares, em ordem decrescente de grau de similaridade.

Sendo assim, o sistema foi desenvolvido para que o cálculo de similaridade seja efetuado, toda vez que um registro satisfizer uma expressão de busca iniciada por "related" - 1º parâmetro do vínculo de similaridade.

A seguir é apresentada a formalização para o cálculo da similaridade.

Seja V o conjunto de palavras.

Seja A : **m** $\longrightarrow$ V o título de uma referência alvo composto por m palavras.

Seja B : **n** $\longrightarrow$ V o título de uma referência similar composto por n palavras.

Onde:

m $= \{1,...,m\}$; e

n $= \{1,...,n\}$.

Seja Ω o subconjunto de palavras comuns a A e B com o acréscimo da palavra x:

$$\Omega = (A(\mathbf{m}) \cup B(\mathbf{n})) + \{x\}.$$

Onde:

$\{x\}$ é um sigleton tal que $\{x\} \cap V = \phi$; e

A(**m**) e B(**n**) são os conjuntos das palavras, respectivamente, dos títulos A e B.

Seja (Ω, $\wp(\Omega)$, P) o espaço de probabilidade uniforme ($P(\{\omega\}) = \frac{1}{|\Omega|}$, para todo $\omega \in \Omega$), onde $|\Omega|$ é a cardinalidade de Ω.

Sejam X e Y duas variáveis aleatórias relativas a este espaço, definidas por:

$$X: \omega \in \Omega \longrightarrow X(\omega) = \begin{cases} 0 \ se \ \omega = x \\ |A^{-1}(\{\omega\})| \ c.c. \end{cases}$$

($X(\omega)$ é a freqüência de ocorrência da palavra ω no título A);

$$Y: \omega \in \Omega \longrightarrow Y(\omega) = \begin{cases} 0 \ se \ \omega = x \\ |B^{-1}(\{\omega\})| \ c.c. \end{cases}$$

($Y(\omega)$ é a freqüência de ocorrência da palavra ω no título B).

A similaridade entre A e B é definida por:

$$sim(A,B) = \frac{1}{2}\left(1 + cor(X,Y)\right),$$

onde $cor(X,Y)$ é o coeficiente de correlação de Pearson dado por (no caso do espaço uniforme):

$$cor(X,Y) = \frac{\sum_{\Omega}(X(\omega)-\bar{X})(Y(\omega)-\bar{Y})}{\sqrt{\sum_{\Omega}(X(\omega)-\bar{X})^2}\cdot\sqrt{\sum_{\Omega}(Y(\omega)-\bar{Y})^2}},$$

onde:

$$\bar{X} = \frac{1}{|\Omega|}\sum_{\Omega}X(\Omega) \quad \text{e} \quad \bar{Y} = \frac{1}{|\Omega|}\sum_{\Omega}Y(\Omega).$$

Observação: por construção X e Y não são variáveis aleatórias constantes e $\mathrm{Var}(X)$ e $\mathrm{Var}(Y)$ são diferentes de zero.

Exemplo

Sejam A e B os títulos definidos por:

$A(1)$ = a $\qquad$ $B(1)$ = a

$A(2)$ = b $\qquad$ $B(2)$ = b

$A(3)$ = c

Neste caso, $\Omega = \{x, a, b, c\}$

As variáveis aleatórias X e Y correspondentes são:

$X(x) = 0$ $\qquad$ $Y(x) = 0$

$X(a) = 1$ $\qquad$ $Y(a) = 1$

$X(b) = 1$ $\qquad$ $Y(b) = 1$

$X(c) = 1$

As respectivas médias e variâncias são:

$\bar{X} = 0{,}75$ $\qquad$ $\mathrm{Var}(X) = 0{,}25$

$\bar{Y} = 0{,}50$ $\qquad$ $\mathrm{Var}(Y) = 0{,}33$

Assim a correlação e a similaridade entre X e Y são:

$cor(X,Y) = 0{,}577$ e $sim(X,Y) = 0{,}79.$

No título não é comum ter uma mesma palavra repetida, portanto geralmente as variáveis aleatórias serão formadas por uma seqüencia de números zero e um. Vale ressaltar que o modo como o cálculo de similaridade está sendo efetuado, permite que este mesmo código também seja utilizado e testado para avaliar a similaridade entre outros campos de

metadados, como por exemplo: entre resumos. O resumo pode apresentar com mais freqüência uma mesma palavra repetida, mas esta metodologia garante de antemão, que a alta freqüência de ocorrência de uma palavra será considerada no cálculo da similaridade.

No cálculo da correlação, considerou-se a probabilidade uniforme. Seria propício melhorar o ajuste dos graus de similaridade atribuindo pesos às palavras consideradas como mais relevantes, mas para isso seria preciso que o sistema atual possuísse recursos que garantissem a correta grafia das palavras, o que o sistema ainda não dispõe. Se por exemplo, palavras do alvo consideradas como relevantes estivessem escritas incorretamente no título da referência similar e recebessem um peso maior, isso resultaria em um irreal baixo grau de similaridade. Sendo assim, para adotar uma probabilidade não uniforme, é preciso dispor de um dicionário de palavras associado ao sistema. Para calcular a similaridade entre palavras, o teste em si não seria complexo, pois demandaria apenas alguns ajustes no sistema atual. Basicamente a diferença seria o fato da ordem das palavras no título não alterar o seu sentido, enquanto a ordem das letras na palavra é fundamental.

Portanto, enquanto o cálculo da similaridade entre os títulos não é sensível à ordem das palavras (fato que levou a detecção de um número razoável de duplicações na análise do grupo controlado de dados bibliográficos), o cálculo da similaridade entre as palavras deve ser sensível à ordem das letras e para isso seria preciso apenas fazer alguns ajustes para que o *domain* registrasse seqüencias de letras e não apenas letras isoladas. Neste experimento, a opção pela probabilidade uniforme no cálculo da correlação mostrou resultados bem favoráveis, pois a relevância das palavras foi prevista na fase anterior ao cálculo da similaridade, durante a busca pelos similares, que contou com um módulo para o pré-processamento dos dados (eliminando palavras sem significado expressivo), combinação de palavras relevantes (selecionadas a partir de outros campos de metadados) e suas inflexões em número.

O algoritmo abaixo resume o processo de cálculo da similaridade, onde `xList` e `yList` são duas listas representando as variáveis aleatórias *X* e *Y*:

```
1.  list ← CONCATENATE(xList, yList)
2.  domain ← TURNUNIQUE(list)
3.  If LENGTH(domain) = 1 Then
4.      Return 1
5.  xList2 ← 0
6.  yList2 ← 0
7.  Foreach item in domain do
8.      nx ← number of occurrences of item in xList
9.      xList2 ← insert nx at the end of xList2
10.     ny ← number of occurrences of item in yList
11.     yList2 ← insert ny at the end of yList2
12. similarity ← (1+CORRELATION(xList2, yList2))/2
13. Return similarity
```

Os valores dos graus de similaridade calculados pelo sistema, não são armazenados no banco de dados, mas calculados em cada acesso a um vínculo de similaridade. Portanto, assim como o processo de montagem destes vínculos, o cálculo de similaridade também é dinâmico, apresentando a mesma vantagem da atualização dos dados em tempo real, i.e., os graus de similaridade e sua ordenação.

A partir dos testes realizados, observou-se que a interface dinâmica dos similares apresentou um rápido retorno. Atualmente, a maior parte dos acessos aos vínculos de similaridade (85% dos vínculos testados) resultou em menos de 10 referências similares, e levou em média menos de dois segundos para sua exibição na tela do cliente. Nos casos em que o acesso ao vínculo de similaridade resulta em mais de dez referências, a interface de retorno da busca apresenta inicialmente às dez referências mais similares ao alvo (com a especificação de seus respectivos graus de similaridade) e um botão com a opção de acesso a todas as referências encontradas. Para que o sistema de recomendação estivesse condizente com a sua função de sugerir o acesso às referências mais similares, a interface padrão de retorno da busca da UR*Lib* não foi adotada, pois para expressões de busca acima de dez referências, o padrão é a apresentação inicial das dez referências mais recentes e não as mais similares. Portanto, foi introduzida uma nova interface de retorno para as buscas disparadas a partir dos vínculos de similaridade.

As submissões e importações de referências na Biblioteca Digital são crescentes, o que indica que nos próximos anos pode haver um aumento considerável na relação de similaridade entre as referências. O ideal é manter o padrão atual, uma média de 10 referências similares acessadas a partir dos vínculos de similaridade. Com este intuito foi criado o 5º PARÂMETRO do vínculo de similaridade. No algoritmo onde consta este parâmetro é definido o número de palavras relevantes e o seu número de combinações. Ao ajustá-los, por exemplo, diminuindo o número de combinações ou aumentando o número de palavras relevantes, automaticamente todos os resultados de busca destes vínculos tornam-se mais restritos, diminuindo o número de resultados similares. Atualmente este ajuste é manual, mas este processo também poderia ser automático, ou seja, ao programar buscas por similares de forma periódica e aleatória, o próprio sistema calcularia o melhor desempenho a partir dos resultados dos testes e efetuaria o auto-ajuste.

4.4.2.1) Propriedade do cálculo de similaridade

A similaridade de A em relação a B é a mesma de B em relação a A. Sendo assim, pode-se afirmar que o cálculo da similaridade possui a propriedade comutativa.

4.5) Estudo estatístico do comportamento das regras de decisões usando a medida de similaridade selecionada

A sensibilidade de cada parâmetro do vínculo de similaridade foi testada utilizando o Sistema de Recomendação para as referências do grupo controlado de dados bibliográficos. Num primeiro momento, a partir da análise da lista de referências obtidas pelos vínculos de similaridade, eram realizados pequenos ajustes nestes parâmetros ou no próprio Sistema de Recomendação.

Ao concluir estes ajustes, foi realizada uma análise mais detalhada, para a definição do grau crítico de similaridade, cujo valor corresponderia ao limiar entre uma referência ser considerada duplicada ou apenas semelhante à referência alvo. O conceito do grau crítico de similaridade é apresentado na **Figura 4**:

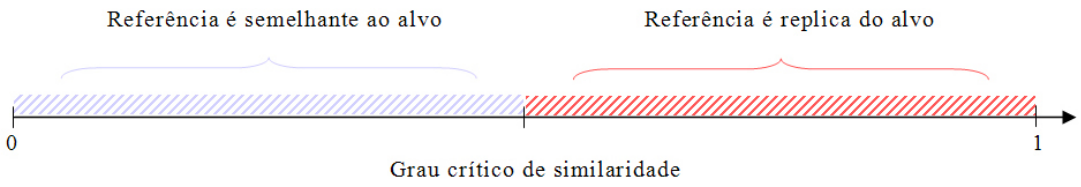

Figura 4 – Grau crítico de similaridade conceitualmente.

Nesta fase, foram avaliados os similares de algumas referências ocultas nos anos de 2008 e 2009. Ao acessar o vínculo de similaridade de uma referência oculta, o sistema de recomendação retornava os similares existentes no acervo, considerando tanto referências ocultas, como visíveis. Nesta análise, foram consideradas as referências similares ocultas e visíveis, mas apenas na mesma categoria de documento da referência alvo, por exemplo: se a referência alvo era da categoria artigo em revista, foram analisados apenas os similares desta categoria. Em seguida, foram registrados os seguintes dados em uma planilha:

Identificador da referência alvo

- Título

Identificador de cada referência similar

- Grau de similaridade
- Estado

  valor 1: duplicado

  valor 0: não duplicado

- Título

A **Figura 5** apresenta a planilha com um exemplo do registro dos dados de uma referência oculta e seus respectivos similares.

ID ALVO	ID SIMILAR	GRAU	STATUS	TITULO
*/12.22.15.56				The turbidity behavior in an Amazon floodplain
	*/12.22.16.13	1,00	1	The turbidity behavior in an Amazon floodplain
	*/12.22.15.29	1,00	1	The turbidity behavior in an Amazon floodplain
	*/06.25.15.56	0,50	1	A contribution to understanding the turbidity behaviour in an Amazon floodplain
	*/10.02.16.48	0,40	0	Evaluation of Spectral Unmixing Algorithm to Modelling the Turbidity Distribution in the Curuai Floodplain (Pará State, Brazilian Amazon)
	*/12.04.13.51	0,39	0	Turbidity in the Amazon Floodplain Assessed Through a Spatial Regression Model Applied to Fraction Images Derived From MODIS/Terra
	*/12.17.01.44	0,39	0	Turbidity in the Amazon floodplain assessed through a spatial regression model applied to fraction images derived from MODIS/Terra

Figura 5 - Planilha para comparação dos dados da referência alvo em relação aos seus similares.

Na figura 5, pode-se observar a existência de similares contendo exatamente o mesmo título. Isto ocorre porque uma mesma referência pode constar no CV Lattes de vários autores e diferentes fontes de uma mesma informação não necessariamente apresentam exatamente o mesmo cadastro. Por exemplo, ambigüidades na identificação de autoria de referências bibliográficas são comuns quando uma Biblioteca Digital recebe informações de diferentes fontes (OLIVEIRA, 2005). No sistema de importação da Plataforma Lattes para a UR*Lib*, estas ambigüidades podem gerar diferentes chaves de citação e conseqüentemente resultar em sucessivas importações de uma mesma referência.

A definição de um valor para o grau crítico de similaridade permitiria incluir facilmente o recurso do cálculo de similaridade durante as importações. No entanto, as análises dos dados demonstraram a existência de uma faixa crítica de similaridade, na qual foi observada a ocorrência tanto de referências duplicadas, como semelhantes. A faixa crítica de similaridade observada durante a análise das referências ocultas é apresentada na **Figura 6**.

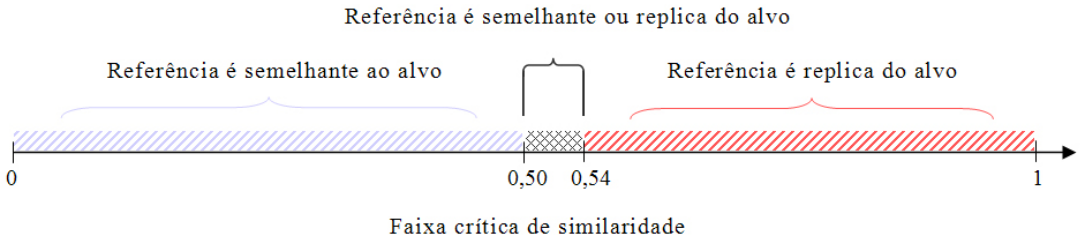

Figura 6 - Faixa crítica de similaridade.

Conforme a **Figura 6**, esta faixa crítica abrangia graus de similaridade com valores entre 0,50 e 0,54. Sendo assim, para graus de similaridade nesta faixa, foram acrescentados na planilha os valores de outros campos de metadados da referência alvo e similar, como por exemplo: autoria, ano, nome da revista, ISSN, volume e página. O intuito foi identificar outros campos de metadados que pudessem auxiliar nas regras de decisão pela aprovação ou reprovação de uma importação, caso o seu grau de similaridade em relação a alguma referência do acervo estivesse dentro desta faixa crítica de similaridade. No entanto, o resultado das análises não indicou um campo de metadado capaz de servir como critério comparativo para assegurar a correta aprovação ou reprovação de uma importação.

A partir das análises, foi possível detectar alguns motivos que tornam comuns as duplicações. Entre os quais, vale ressaltar os casos de referências cadastradas incorretamente no Lattes como publicadas, enquanto deveriam ter sido cadastradas como submetidas. Este equívoco acarreta a importação da versão preliminar da referência. O documento, em geral, sofre mudanças significativas em diferentes campos até a sua publicação efetiva e neste período, foram observadas alterações em vários campos, como por exemplo: autoria, título, ano, nome da revista, ISSN, volume e página. Com a publicação do artigo, os outros autores do documento em questão, cadastram corretamente em seus currículos a obra publicada e conseqüentemente ocorre outra importação da mesma referência, gerando a duplicação no acervo institucional. Nos estudos de casos, as mudanças ocorridas no título, deixaram o duplicado na faixa crítica de similaridade e a consulta aos outros campos da referência alvo e similar não foram conclusivas.

Nesta fase, a revisão e a experiência dos bibliotecários foram fundamentais para a correta interpretação dos casos de similaridade. A partir das análises, optou-se por não incluir o cálculo de similaridade durante as importações, pois o processo de automação deveria evitar as duplicações, mas sem prejuízo às importações. Sem a possibilidade de definir uma regra, as referências semelhantes, na faixa crítica de similaridade, deixariam de ser importadas, prejudicando desnecessariamente o processo de importação.

4.6) Similar Search UR*Lib* Accessory - a proposta de uma interface web para gerenciar as referências similares na Biblioteca Digital.

Como a inclusão do cálculo de similaridade durante as importações poderia inviabilizar a importação de um número imprevisível de referências, esta etapa foi substituída pela proposta de uma interface web - Similar Search UR*Lib* Accessory (SIMILAR-SEARCH) - oferecendo aos bibliotecários, o gerenciamento dos casos de similaridade no acervo.

A nova proposta, ao contrário da anterior, não traria prejuízos às importações, pois o cálculo de similaridade não seria efetuado durante as importações, mas após a conclusão deste processo. A partir de um relatório gerado pelo sistema, os bibliotecários teriam o controle de todas as referências importadas cujo grau de similaridade fosse igual ou superior a 0,50. Neste novo enfoque, a prioridade foi a definição do menor grau de similaridade em que se constatou pelo menos um caso de duplicação. Ao obter este valor, conseqüentemente, grande parte das referências na faixa crítica de similaridade estaria incluída no relatório para avaliação dos bibliotecários.

O fluxo de processos do SIMILAR-SEARCH é descrito a seguir:

⇒ O bibliotecário, com permissão de acesso, seleciona um grupo de dados bibliográficos, especificando o ano e a categoria das referências importadas, que serão analisadas.

⇒ Cada referência importada é avaliada a partir do sistema de recomendação, como uma referência alvo. Os respectivos similares de cada alvo são, caso a caso, listados na interface do SIMILAR-SEARCH.

⇒ Ao apresentar um alvo na interface, ao lado de cada uma das referências similares, o bibliotecário tem a opção de identificar o estado de "Duplicado" de cada referência similar. No momento em que o bibliotecário seleciona este estado, o SIMILAR-SEARCH registra nos metadados desta referência, uma tag com o estado escolhido e o identificador da referência alvo, como por exemplo: <DUPLICATE: id1>. Este procedimento preserva o histórico do conteúdo revisado, evitando que o sistema exiba novamente um duplicado analisado anteriormente, além de assegurar um vasto material para futuros estudos sobre a similaridade.

⇒ Ao concluir a análise de cada caso, o relatório é salvo e o sistema exibe o próximo caso para ser avaliado. Neste contexto, o bibliotecário não precisa revisar todas as referências importadas em uma única etapa, ou seja, se precisar interromper a análise, todos os casos revisados até aquele momento estarão salvos e ao retornar ao sistema, a análise recomeça do ponto em que houve a interrupção.

⇒ Quando a revisão das referências importadas estiver concluída:

- O administrador de sistemas usando a interface do SIMILAR-SEARCH gera a lista de referências duplicadas, que foram selecionadas pelos bibliotecários para serem ocultas das interfaces exibidas ao usuário final.

- O bibliotecário gera pela interface web do SIMILAR-SEARCH um histórico de todas as referências analisadas até o momento.

A listagem caso a caso e o registro das análises nos metadados foram dois fatores fundamentais para solucionar várias questões. Por exemplo: supondo que o sistema tenha importado duas referências identificadas como: R1 e R2; e ao avaliá-las como dois alvos, o sistema detectasse os seguintes similares (com grau crítico maior ou igual a 0,50) apresentados na **Figura 7**:

Alvo	Similares com G ≥ 0,50
R1	R2
R2	R1

Figura 7 – Os alvos R1 e R2 apresentando respectivamente os similares R2 e R1.

Observando este exemplo, é possível comprovar que caso o sistema se limitasse a listar todos os alvos e seus similares, sempre que duas referências importadas apresentassem entre si uma similaridade acima do grau crítico, o bibliotecário teria o retrabalho de avaliar uma mesma relação de similaridade. Isso ocorre, pois como descrito no item 4.4.2.1 deste relatório, o cálculo de similaridade é uma operação comutativa, ou seja, a similaridade de R1 em relação a R2 é a mesma que R2 em relação a R1. No SIMILAR-SEARCH a apresentação é caso a caso, sendo assim, uma vez que a análise de R1 em relação a R2 resultou, por exemplo, na seleção de R2 como um duplicado do alvo R1, os metadados de R2 terão a identificação da tag "duplicado". Como o sistema descarta a análise das referências com esta tag, conseqüentemente não haverá a análise de R2 em relação a R1.

O SIMILAR-SEARCH deve ser executado enquanto forem detectados casos de similaridade, pois a convergência do algoritmo não ocorre em um único passo, ou seja, não é um processo idempotente.

A aplicação do SIMILAR-SEARCH para todas as referências do grupo controlado de dados bibliográficos apresentou resultados satisfatórios:

- em 95% dos casos, a nova proposta indicou corretamente o duplicado e descartou a análise de referências semelhantes (não duplicadas);

- em 3% dos casos, as referências semelhantes foram sugeridas para análise da similaridade em relação ao alvo;

- em 2% dos casos, o sistema falhou descartando da análise referências duplicadas.

5) Resultados obtidos em função do plano de trabalho proposto

Desenvolvimento de um programa para efetuar o cálculo de similaridade, a partir de uma medida de similaridade com base na correlação de *Pearson*.

Desenvolvimento em TCL de um Sistema de Recomendação para a Biblioteca Digital, utilizado neste projeto com o intuito de testar o comportamento do grupo controlado de dados bibliográficos, para a definição de um valor arbitrário para o grau crítico de similaridade.

Proposta de uma interface web para gerenciar as referências similares na Biblioteca Digital - Similar Search UR*Lib* Accessory. A nova proposta apresentou resultados satisfatórios quando aplicada para o grupo controlado de dados bibliográficos, tendo um desempenho favorável em 98% dos casos.

6) Publicações científicas durante o período da bolsa

RIBEIRO, M. L.; BANON, G. J. F.; BANON, L. C. Repositório digital dos anais do SBSR do INPE. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 16. (SNBU) - SEMINÁRIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS-BRASIL, 2. (SIBD-B), 2010, Rio de Janeiro. **Anais...** 2010. On-line. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2010/ 10.22.11.57>.

RIBEIRO, M. L.; BANON, G. J. F.; BANON, L. C. **Repositório digital dos anais do SBSR do INPE**. São José dos Campos: INPE, 2010. (INPE ePrint sid.inpe.br/mtc-m19@80/2010/06.30.17.09). Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19@80/2010/ 06.30.17.09>.

7) Conclusões gerais

Este projeto consistiu na pesquisa de soluções para a identificação de réplicas de dados bibliográficos que não possuem identificadores persistentes. Inicialmente, foram realizados estudos para a montagem do grupo controlado de dados bibliográficos. A partir deste estudo, concluiu-se que o título seria o campo primordial para o cálculo de similaridade.

Para o objetivo do projeto em questão, observou-se através das análises dos títulos, a necessidade de adotar uma medida de similaridade cuja abordagem fosse flexível e pudesse não apresentar sensibilidade a ordem das palavras. A partir de uma medida de similaridade com base na correlação de *Pearson*, foi desenvolvido um programa para efetuar o cálculo de similaridade, que associado ao vínculo de similaridade, resultou em um Sistema de Recomendação para a Biblioteca Digital. Este novo recurso foi eficiente para testar o comportamento do grupo controlado de dados bibliográficos e na definição de um valor para o grau crítico de similaridade.

Com as análises realizadas a partir do Sistema de Recomendação concluiu-se que, a aplicação do cálculo de similaridade durante as importações seria um processo prematuro, pois os testes demonstraram a existência de uma faixa crítica de similaridade, na qual o sistema impediria a importação de referências não duplicadas (cujos graus de similaridade fossem iguais ou superiores ao crítico), um indício de que a automação poderia trazer prejuízos às importações da Plataforma Lattes para a UR*Lib*.

A etapa de inclusão do cálculo de similaridade durante as importações foi substituída pela proposta de uma interface web - Similar Search UR*Lib* Accessory (SIMILAR-SEARCH) – com o objetivo de oferecer aos bibliotecários, o gerenciamento dos casos de similaridade no acervo.

A aplicação do SIMILAR-SEARCH para todas as referências do grupo controlado de dados bibliográficos apresentou resultados satisfatórios, tendo um desempenho favorável em 98% dos casos.

Referências Bibliográficas

BANON, G. J. F. **Biblioteca Digital da Memória Técnico-Científica do INPE.** São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2006-03-11. (INPE ePrint dpi.inpe.br/banon-pc2@1905/2005/12.07.19.19). Disponível em: <http://ePrint.sid.inpe.br:80/rep-/dpi.inpe.br/banon-pc2@1905/2005/12.07.19.19>. Acesso em: 10 jan. 2009.

CORMEN T.H.; LEISERSON C.E.; RIVEST R.L; STEIN C. **Introduction to algorithms**. 2 ed. Cambridge: MIT Press; 2001. ISBN 0262032937

NEVEU, J. **Bases Mathématiques du Calcul des Probabilités.** Paris: Masson. 1970.

OLIVEIRA, J. W. A.; LAENDER, A. H. F.; GONÇALVES, M. A. Remoção de ambigüidades na identificação de autoria de objetos bibliográficos. In: SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS, 20. (SBBD), 2005, Uberlândia. Anais... Uberlândia: UFU, p. 205-219. On-line. ISBN 85-7669-029-2. Disponível em: <http://www.sbbd-sbes2005.ufu.br/arquivos/artigo-14-OliveiraLaender.pdf>. Acesso em: 21 jan. 2009.

PAPOULIS, A. **Probability & statistics**. Englewood Cliffs, NJ: Prentice-Hall, 1990. 454 p. ISBN 013711698-5.

TRISKA, R.; CAFÉ, L. Arquivos abertos: subprojeto da Biblioteca Digital Brasileira. In: **Ciência da Informação**, Brasília, 30(3):92-96, set./dez.2001 Disponível em: <http://www.scielo.br/pdf/ ci/v30n3/7291.pdf >. Acesso em: 21 ago. 2008.

São José dos Campos, 26 de Novembro de 2010

_____

Lise Christine Banon

_____

Dr. José Carlos Neves Epiphanio

Orientador PCI/OBT