Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Expansão de Consultas com Realimentação e Pseudo Realimentação de Relevantes em um Sistema que utiliza o Modelo TR+ para Indexar e Recuperar Documentos

Thyago Bohrer Borges

Orientador: Profa. Dra. Vera Lúcia Strube de Lima

Porto Alegre
2009

Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Expansão de Consultas com Realimentação e Pseudo Realimentação de Relevantes em um Sistema que utiliza o Modelo TR+ para Indexar e Recuperar Documentos

Thyago Bohrer Borges

Dissertação apresentada como requisito parcial à obtenção do grau de mestre em Ciência da Computação

Orientador: Profa. Dra. Vera Lúcia Strube de Lima

Porto Alegre
2009

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "**Expansão de Consultas com Realimentação e Pseudo Realimentação de Relevantes em um Sistema que utiliza o Modelo TR+ para Indexar e Recuperar Documentos**", apresentada por Thyago Bohrer Borges, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional, aprovada em 26/01/09 pela Comissão Examinadora:

Vera Lúcia Strube de Lima

Profa. Dra. Vera Lúcia Strube de Lima– Orientadora — PPGCC/PUCRS

Prof. Dr. Marcelo Blois Ribeiro - — PPGCC/PUCRS

Prof. Dr. Marco Antonio Insaurriaga Gonzalez – — FACIN/PUCRS

Profa. Dra. Viviane Moreira Orengo – — UFRGS

Homologada em 15/09/09, conforme Ata No. 16/09 pela Comissão Coordenadora.

Prof. Dr. Fernando Gehm Moraes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

Dados Internacionais de Catalogação na Publicação (CIP)

B732e Borges, Thyago Bohrer
Expansão de consultas com realimentação e pseudo realimentação de relevantes em um sistema que utiliza o modelo TR+ para indexar e recuperar documentos / Thyago Bohrer Borges. – Porto Alegre, 2009.
169 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientadora: Profa. Dra. Vera Lúcia Strube de Lima.

1. Informática. 2. Sistemas de Recuperação da Informação. 3. Processamento da Linguagem Natural. 4. Linguística Computacional. I. Lima, Vera Lúcia Strube de. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Dedico este trabalho a todos aqueles que torceram por mim.

Agradecimentos

Agradeço a Deus por ter me abençoado com saúde, paz e serenidade nos momentos difíceis durante o desenvolvimento deste trabalho. Agradeço aos meus pais, Jorge e Gladis, que sempre acreditaram em mim, me apoiando e não permitindo que eu esqueça do meu potencial. Minhas vitórias são conseqüência de seus esforços em me tornar um ser humano íntegro e lutador. Agradeço a Francieli Carniel, que sempre torceu por mim, me apoiando e incentivando nos momentos difíceis, mostrando paciência e compreensão. Agradeço a Profa. Dra. Vera Lúcia Strube de Lima, que me deu a oportunidade e o suporte necessários para a realização e conclusão deste trabalho, sempre disposta a sanar minhas dúvidas me orientando para o melhor caminho a ser seguido durante o desenvolvido da dissertação. Agradeço ao Prof. Dr. Marco Gonzalez pela generosidade e compromisso com o compartilhamento do conhecimento, por ter me dado o suporte necessário para o entendimento acerca do Modelo TR+. Agradeço a Pontifícia Universidade Católica do Rio Grande do Sul e ao Centro de Desenvolvimento e Pesquisa - DELL/PUCRS, que possibilitaram este período de estudos. Agradeço aos novos amigos que conheci durante este período e que comigo participaram desta etapa, amigos do CDPe e PLN. Agradeço aos velhos e bons amigos Tiago Silva e Tiago Primo, companheiros para todas as horas, são amizades assim que facilitam as conquistas. Agradeço ao Prof. Dr. Stanley Loh que me ensinou o gosto pela pesquisa científica.

”No meio da dificuldade encontra-se a oportunidade”

Albert Einstein

”Não podemos parar...”

Irai Borges

Resumo

Este trabalho apresenta e discute os resultados obtidos com a aplicação das técnicas de expansão de consulta denominadas Pseudo Realimentação de Relevantes (PRR) e Realimentação de Relevantes (RR) em um Sistema de Recuperação de Informação (SRI) que utiliza o modelo de recuperação de informação denominado TR+. TR+ é um modelo de recuperação de informação que emprega, além de termos, Relações Lexicais Binárias (RLB) presentes nos textos e nas consultas, para indexar e recuperar documentos textuais em língua portuguesa. A aplicação das técnicas de expansão de consultas PRR e RR têm como objetivo melhorar os resultados obtidos pelo usuário que realiza uma consulta. As duas técnicas se diferenciam quanto à participação do usuário: enquanto a RR utiliza o julgamento do usuário na definição de quais documentos recuperados pela consulta original fornecerão as informações utilizadas na expansão da consulta, a PRR busca eliminar a participação do usuário durante este processo. Os resultados obtidos pelos experimentos, tanto utilizando PRR quanto RR, não superaram os resultados utilizados como *baseline* (Gonzalez, 2005). Ao compararmos entre si os resultados dos experimentos com as técnicas PRR e RR, os experimentos com PRR foram superados pela RR somente em uma rodada. No contexto dessa dissertação podemos concluir que a utilização de RLBs ao invés de usar somente termos, é uma opção mais producente.

Palavras-chave: Expansão de Consultas, Sistemas de Recuperação de Informação, Pseudo Realimentação de Relevantes, Realimentação de Relevantes, Relações Lexicais Binárias, Processamento da Língua Natural.

Abstract

This work presents and debates the results of applying query expansion techniques such as Pseudo Relevance Feedback (PRF) and Relevance Feedback (RF) in an Information Retrieval System (IRS) that uses the information retrieval model TR+. TR+ makes use of terms and Binary Lexical Relationships (BLR) that appear in texts and queries in order to index and retrieve textual documents in Portuguese. The application of the query expansion techniques PRR and RR aims to improve the results provided by the users' queries therefore the documents retrieved are able to fulfill their needs. PRR and RR differ with respect to the users' role: while relevance feedback makes use of the user judgment for defining which documents retrieved by the original query will provide the information for QE, PRF seeks to automate such decision process. The experimental results using PRF and RF did not outperform the baseline results (Gonzalez, 2005). When comparing both techniques, we have noticed PRF was outperformed by RF only once. In the context of this dissertation, we can conclude that the use of BLRs is a more productive option when compared to the use of terms for QE.

Keywords: Query Expansion, Information Retrieval Systems, Pseudo Relevance Feedback, Relevance Feedback, Binary Lexical Relations, Natural Language Processing.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

CLEF	*Cross-Language Evaluation Forum*
EC	Expansão de Consulta
LSI	*Latent Semantic Indexing*
PRR	Pseudo Realimentação de Relevantes
RI	Recuperação de Informação
RLB	Relações Lexicais Binárias
RR	Realimentação de Relevantes
SRI	Sistemas de Recuperação de Informação
TREC	*Text Retrieval Conference*

Sumário

1 Introdução

1.1 Motivação e contexto do trabalho

Para auxiliar os usuários a encontrar, em um repositório com um grande volume de documentos em formato digital, os documentos que necessitam, foram desenvolvidos os Sistemas de Recuperação de Informação (SRIs) (Baeza-Yates & Ribeiro-Netto, 1999). SRIs que trabalham com documentos textuais tem como objetivo principal atender consultas realizadas por usuários através da indexação, classificação e busca de documentos (Salton & MacGill, 1983). Entretanto formular uma consulta correta através de palavras-chave, que possibilite a um SRI retornar ao usuário as informações que ele necessita, pode não ser uma tarefa simples. Segundo Baeza-Yates e Ribeiro-Netto (1999), a identificação da real necessidade do usuário no momento em que ele busca alguma informação ou documento é um processo muito complexo e pode ser a diferença entre uma recuperação producente e uma recuperação de informações que não atenda as necessidades do usuário.

Apesar de auxiliarem a recuperação de informações, os SRIs dependem da capacidade do usuário de formular eficientemente uma consulta, de maneira que o sistema possa ”interpretar” o que o usuário deseja obter como resposta, no momento da recuperação das informações.

Uma alternativa para ajudar ao usuário na formulação da consulta e, por conseqüência, aumentar a eficiência de um SRI, é a utilização de uma técnica denominada Expansão de Consulta (EC). EC, utilizando novos termos semanticamente relacionados aos já presentes na consulta inicial, é uma técnica tradicional na recuperação de informações (Monz, 2003).

Estudos realizados por Spink et al. (2001), embora não muito recentes, apontam que 52% das consultas realizadas em um sistema de recuperação de informação são refeitas. Outro estudo, apresentado em abril de 2006 pela iProspect[1], empresa pioneira no desenvolvimento de SRIs para marketing, mostra que 82% dos usuários refazem suas consultas acrescentando mais termos, realizando com isso uma expansão da consulta original com novos termos, buscando um meio de encontrar as informações que melhor satisfaçam as suas necessidades.

A utilização de técnicas de expansão de consulta, em conjunto com técnicas de processamento da língua natural (PLN), constitui-se nos dias de hoje uma alternativa viável para se aprimorar o resultado da recuperação, em um sistema de recuperação de informação.

Ao longo desse trabalho foram estudados os processos e analisados os resultados da aplicação das técnicas de expansão de consulta Pseudo Realimentação de Relevantes (PRR) e Reali-

[1] Disponível em http://www.iprospect.com/media/

mentação de Relevantes (RR) em um sistema de recuperação de informação que usa o Modelo TR+ (Gonzalez, 2005) para indexar e recuperar documentos textuais em língua portuguesa.

O Modelo TR+ utiliza métodos estatísticos e lingüísticos para indexar e recuperar os documentos, e apresenta características apropriadas para a representação e recuperação textual. Propõe a utilização de tratamento idêntico para os textos dos documentos e para as consultas formuladas através de palavras-chave, e reformuladas com a inclusão de operadores booleanos antes do processo de recuperação.

O Modelo TR+ indexa e recupera utilizando, tanto sobre a formulação da consulta quanto na indexação dos documentos, descritores de conceitos que incluem termos simples e compostos (Gonzalez et al., 2006a). O primeiro passo é o pré-processamento do texto, onde são utilizados métodos de tokenização e etiquetagem morfológica. Após, são realizadas a nominalização e a captura das relações lexicais binárias (RLBs).

Nominalização, no Modelo TR+, é o processo de transformação de adjetivos, verbos e advérbios em substantivos (Gonzalez et al., 2005). Com a nominalização são definidos os termos simples que constituirão os descritores. As RLBs (Gonzalez et al., 2006b) constituem os termos compostos e completam a descrição dos conceitos presentes nos documentos. RLBs (dos tipos classificação, restrição e associação) são relacionamentos entre termos nominalizados, que capturam mecanismos de coesão frásica (Gonzalez, 2005).

Os descritores (termos e RLBs) têm seus pesos calculados através do conceito de evidência. Neste cálculo é considerada a freqüência de ocorrência dos termos e, também, o número de relações que há entre eles.

A seguir na Seção 1.2 introduziremos os objetivos do presente trabalho.

1.2 Hipótese e objetivos

Baeza-Yates e Ribeiro-Netto (1999) apresentam várias técnicas de expansão de consulta que buscam a formulação de uma consulta mais eficiente, em um sistema de recuperação de informação. Entretanto, nenhuma dessas técnicas havia sido utilizada em conjunto com o Modelo TR+ em um sistema de recuperação de informação. Deste modo, a questão de pesquisa ganha corpo da seguinte forma:

"A aplicação de expansão de consulta em um sistema de recuperação de informação que usa o Modelo TR+ pode representar um ganho de precisão ou abrangência na recuperação dos documentos?"

Partindo-se desta questão de pesquisa, tem-se como hipótese que a utilização de técnicas de expansão de consulta em conjunto com o Modelo TR+ pode auxiliar na recuperação de informações, aumentando a precisão das informações recuperadas pelo sistema de recuperação de informação.

Esse desempenho pode ser analisado, por exemplo: (i) pela comparação da utilização dos

termos nominalizados (substantivos concretos e abstratos); (ii) pelo número de RLBs utilizadas na expansão da consulta; (iii) pelo número de termos utilizados na expansão da consulta. O desempenho das técnicas de EC pode ser avaliado por: (i) pelo cálculo da precisão, (ii) abrangência e (iii) MAP das informações recuperadas pelo sistema. No contexto dessa dissertação, no que tange a avaliação dos experimentos, foram realizadas as análises: (i) número de RLBs utilizadas para a EC; (ii) tipo de RLBs utilizadas para a EC, (iii) número de termos utilizados para a EC. E para a avaliação foram realizados cálculo das métricas: (i) precisão, (ii) abrangência e (iii) MAP.

A dissertação tem como objetivo geral estudar os efeitos resultantes da aplicação de expansão de consulta em conjunto com o Modelo TR+ em um sistema de recuperação de informação. Espera-se que a EC auxilie o usuário, possibilitando que o sistema obtenha maior eficiência no atendimento de suas expectativas. A técnica empregada vem agregar suas características à estratégia não clássica de representação de documentos e consultas já utilizada pelo modelo.

1.3 Organização do texto da dissertação

O texto da dissertação está organizado em 7 capítulos, seguido de referências e anexos.

O Capítulo 2, apresenta os modelos tradicionais de recuperação de informação (Seção 2.1). Modelos como o Booleano (Seção 2.1.1), o Modelo Espaço Vetorial (Seção 2.1.2) e o Probabilístico (Seção 2.1.3) são abordados, com suas características.

O Capítulo 3, apresenta um apanhado geral de técnicas tradicionais de EC entendidas por nós como importantes para este trabalho. Neste capítulo descrevemos técnicas com e sem a participação do usuário para a EC.

No Capítulo 4 apresentamos as características do Modelo TR+, seu funcionamento e suas peculiaridades.

No Capítulo 5 apresentamos alguns trabalhos recentes que utilizam EC em seus estudos e que serviram, de alguma forma, para nortear o trabalho desenvolvido nesta dissertação.

No Capítulo 6 apresentamos os experimentos realizados para a validação da proposta da dissertação assim como os resultados obtidos em cada caso.

No Capítulo 7 finalizamos a dissertação com as considerações sobre o trabalho desenvolvido, apontando contribuições, limitações e algumas propostas para sua continuidade.

2 Recuperação de Informação

Segundo Baeza-Yates e Ribeiro-Netto (1999) Recuperação de Informação (RI) tem o objetivo de representar, armazenar, organizar e acessar alguma informação. A representação e organização deve possibilitar, ao usuário, acesso rápido e fácil às informações de seu interesse. Manning et al. (2008) definem o papel de um SRI como sendo de encontrar material (geralmente documentos), de uma natureza não estruturada (geralmente texto), que satisfaça a necessidade de uma informação dentro de grandes coleções (geralmente armazenadas em computadores). Entretanto a identificação da real necessidade do usuário no momento em que ele busca alguma informação não é uma tarefa trivial.

Para tornar mais fácil ao usuário encontrar informações desejadas em grandes repositórios de informações, foram desenvolvidos os Sistemas de Recuperação de Informação (SRI). Um SRI, é um sistema de computador capaz de armazenar, recuperar e manter informações (Kowalski, 2000). Com a criação dos SRIs pôde-se então minimizar o esforço do usuário ao procurar a informação desejada.

Dentre as informações ou itens recuperados por um SRI, estão os documentos textuais. SRIs que se preocupam em recuperar documentos textuais têm a finalidade de encontrar documentos que possam conter a resposta para alguma questão que o usuário necessite responder e não encontrar a resposta em si (Monz, 2003). Na Figura 1 apresentamos o processo de recuperação de informação proveniente de Orengo (2004).

Na Figura 2 apresentada por Chen e Gey (2004), podemos observar os componentes que constituem a arquitetura de um SRI.

A base para qualquer sistema de recuperaçao de informação são os chamados modelos clássicos de recuperação de informação. Na Seção 2.1 serão apresentados os modelos clássicos de SRIs, assim como suas características e seu funcionamento.

2.1 Modelos Clássicos de Sistemas de Recuperação de Informação

Os modelos clássicos de SRIs, consideram que cada documento é representado por uma ou mais palavras-chave que é chamado de termo de índice (do inglês *index terms*), sendo um termo de índice simplesmente uma ou mais palavras que ajudam a encontrar o tema principal do documento (Baeza-Yates & Ribeiro-Netto, 1999). Assim estes termos são utilizados para indexar e sumarizar o conteúdo do documento. Em geral, segundo Baeza-Yates e Ribeiro-Netto

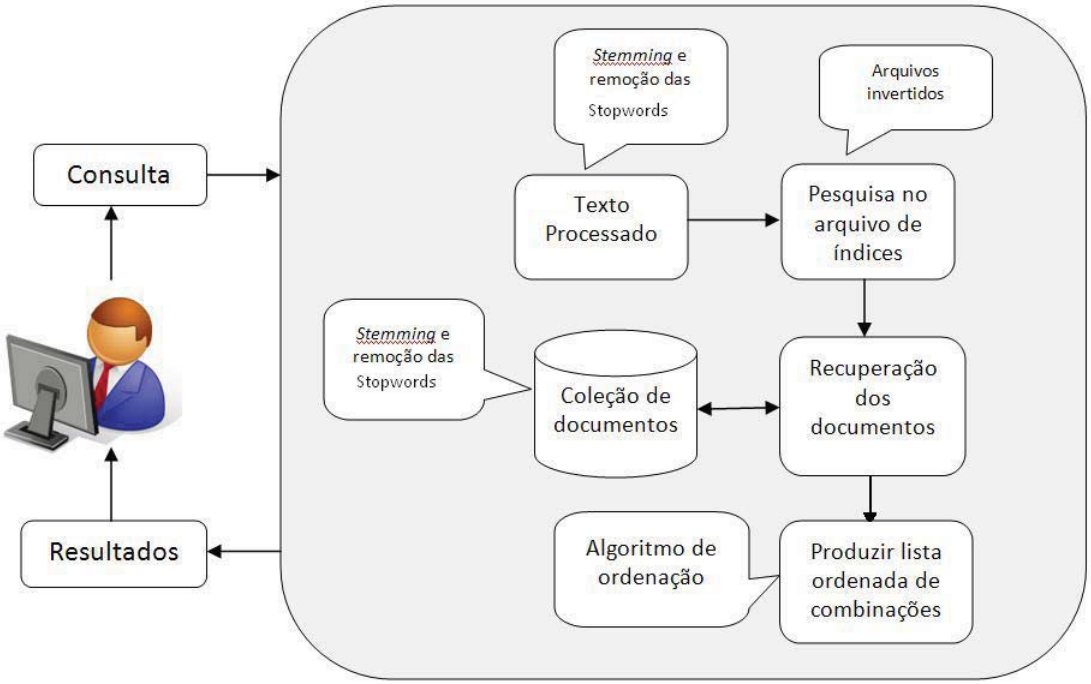

Figura 1 – O Processo de Recuperação de Informação (Orengo, 2004)

(1999) as palavras que são utilizadas como termos de índice são principalmente substantivos. A utilização de substantivos como termos de índice se justifica, pelo fato de que substantivos possuem uma semântica fácil de ser identificada. Adjetivos, advérbios e conectivos são menos úteis como termos de índice porque eles trabalham principalmente como complementos.

Nem todos os termos de um documento podem ser utilizados como um termo de índice, pois não são todos os termos que são capazes de descrever o conteúdo de um documento. De fato alguns termos são menos expressivos que outros, não contribuindo para a identificação do conteúdo do documento a que ele pertence. Decidir a importância de um termo para a indexação e sumarização de um documento não é uma tarefa simples (Baeza-Yates & Ribeiro-Netto, 1999). Apesar desta dificuldade, as propriedades de um termo de índice podem ser medidas, sendo esta medida útil para avaliar o potencial de cada termo para descrever o conteúdo de um documento. Podemos exemplificar da seguinte maneira: em uma coleção de documentos que possua centenas de milhares de documentos, uma palavra que apareça em todos os documentos é completamente inútil como um termo de índice, pois através deste termo não é possível identificar o documento que o usuário deseja recuperar. Por outro lado, seguindo o mesmo exemplo quanto ao número de documentos, palavras que estão presentes somente em poucos documentos são úteis como termos de índice, pois restringem e muito o número de documentos que o usuário possa ter interesse de recuperar (Baeza-Yates & Ribeiro-Netto, 1999). Com isso, podemos identificar quanto a sua relevância através de termos que estejam dentro de um documento.

Para se diferenciar as relevâncias entre os termos de índice quanto a sua capacidade de descrever o conteúdo de um documento, segundo Baeza-Yates e Ribeiro-Netto (1999), é utilizada

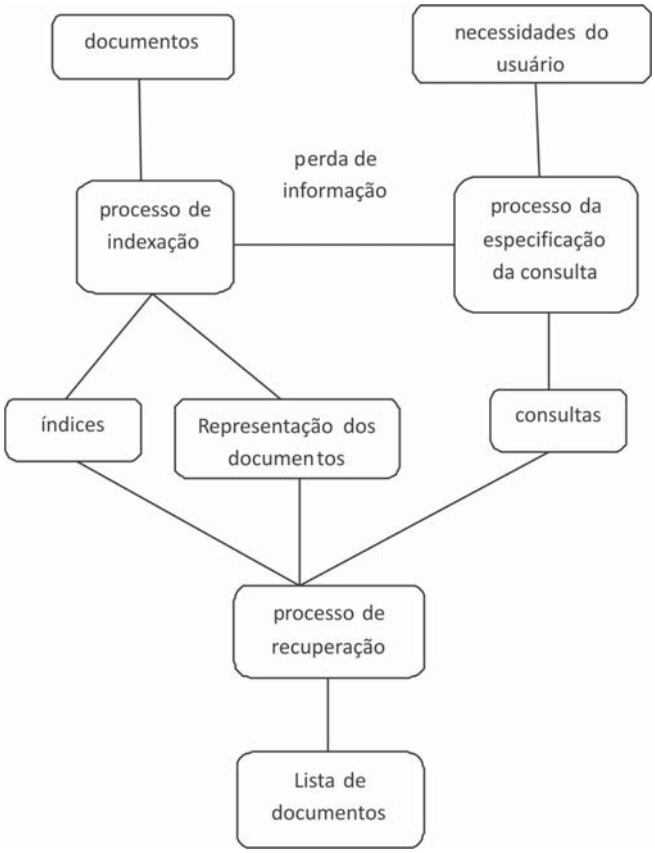

Figura 2 – Arquitetura de um Sistema de Recuperação de Informação (Chen & Gey, 2004)

uma assinatura numérica para cada termo de índice denominada peso.

Baeza-Yates e Ribeiro-Netto (1999) definem o grau (peso) que identifica a relevância de um termo quanto a sua capacidade de representar um documento como: seja k_i um termo de índice, seja d_j um documento, e seja $w_{i,j}$ >= 0 um peso associado ao par k_i,d_j.

Entre os modelos de informação existe uma classificação a qual é denominada taxonomia dos SRIs (Baeza-Yates & Ribeiro-Netto, 1999). Tal taxonomia não será tratada em seu todo, sendo alvo deste estudo os chamados modelos clássicos dos sistemas de informação. Na Figura 3 (Baeza-Yates & Ribeiro-Netto, 1999), é possível ver a taxomonia dos modelos de recuperação de informação.

Apresentamos uma breve revisão teórica para que possamos ingressar com mais profundidade no estudo dos três modelos clássicos dos SRIs propostos na literatura. Os modelos clássicos de recuperação de informação que abordaremos nas subseções 2.1.1, 2.1.2 e 2.1.3, são respectivamente: modelo Booleano, modelo Vetorial e modelo Probabilístico.

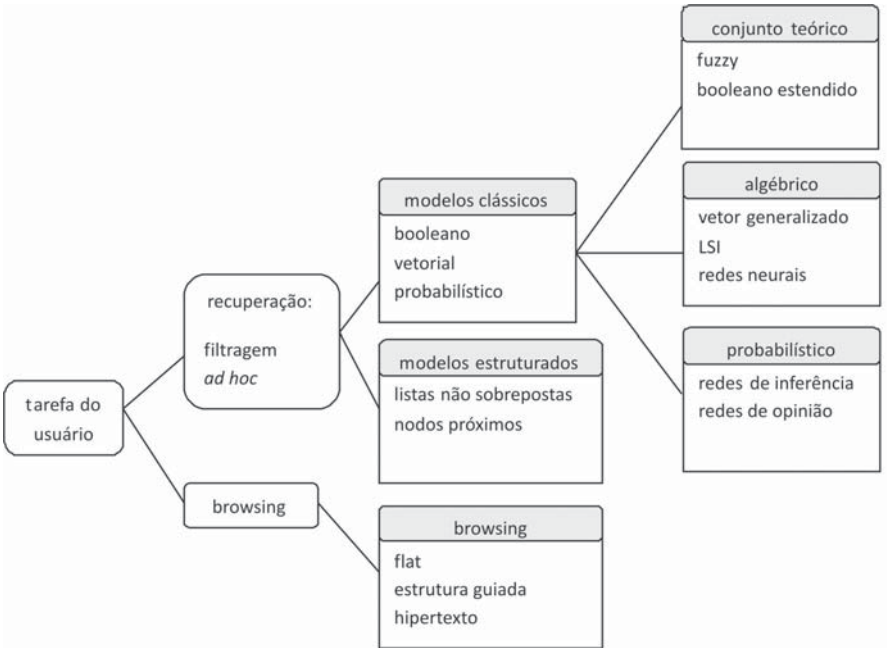

Figura 3 – Taxonomia dos modelos de recuperação de informação (Baeza-Yates & Ribeiro-Netto, 1999)

2.1.1 Modelo Booleano

O modelo Booleano é reconhecido como um modelo de recuperação de informação simples baseado na teoria dos conjuntos e na álgebra booleana (Baeza-Yates & Ribeiro-Netto, 1999). Partindo da possibilidade de se intuir um conceito em um conjunto, o modelo Booleano facilita aos usuários entender o funcionamento de um sistema qualquer de recuperação de informação (SRI). Segundo Manning et al. (2008) o modelo booleano de recuperação de informação possibilita traduzir qualquer consulta escrita por palavras-chave na forma de uma expressão booleana de termos, ou seja, combinando os termos e operadores Booleanos (*AND*, *OR*, *NOT*). As consultas no modelo Booleano são especificadas em uma expressão Booleana a qual possui uma semântica precisa (Baeza-Yates & Ribeiro-Netto, 1999). O modelo Booleano recebe grande atenção por parte de desenvolvedores de sistemas comerciais, isso se deve pela sua inerente simplicidade e seu forte formalismo. Na Figura 4 (Baeza-Yates & Ribeiro-Netto, 1999), podemos ver o modelo Booleano quanto a sua concepção junto a teoria dos conjuntos.

De acordo com Baeza-Yates e Ribeiro-Netto (1999) o modelo Booleano possui sua estratégia de recuperação é baseada em um critério de decisão binária (o documento recuperado é relevante ou não é relevante à consulta efetuada), sem nenhuma noção de classificação, o qual impede que esse modelo tenha um bom desempenho na recuperação de documentos. Segundo Baeza-Yates e Ribeiro-Netto (1999), uma vez que uma expressão Booleana possui uma semântica precisa, a tradução das informações necessárias para uma expressão Booleana é um processo nada trivial. Por isso muitos usuários tem dificuldades em expressar suas necessidades com clareza na hora de realizarem sua consulta no sistema.

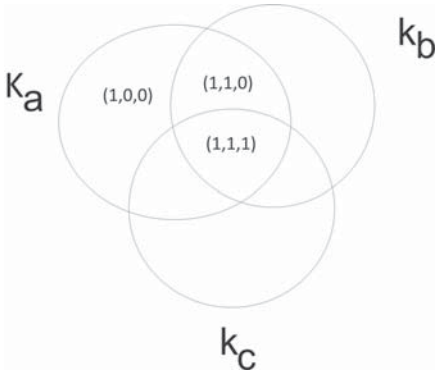

Figura 4 – Conjunção dos três conectivos que compõem uma consulta [q = $k_a \wedge (k_b \vee \neg k_c)$] (Baeza-Yates & Ribeiro-Netto, 1999)

Considera-se a presença ou não dos termos de índice no modelo booleano em um documento. Em conseqüência disto, assume-se que os pesos dos termos de índice são todos binários, onde: $w_{i,j} \in \{0, 1\}$.

Convencionalmente uma consulta *q* é essencialmente composta por termos de índice e pelos três conectivos lógicos: *NOT*; *AND* e *OR* (Baeza-Yates & Ribeiro-Netto, 1999).

Baeza-Yates e Ribeiro-Netto (1999), explicita que o modelo Booleano tenta identificar a não relevância do documento à consulta realizada. O modelo Booleano não possui uma combinação parcial para uma condição da consulta. A maior vantagem do modelo Booleano é seu formalismo claro e sua simplicidade (Salton & Buckley, 1997). A principal desvantagem é a necessidade de se ter uma exata combinação entre a consulta e os documentos a serem recuperados, podendo ocorrer portanto, uma recuperação de um número muito pequeno ou um número muito grande de documentos (Baeza-Yates & Ribeiro-Netto, 1999). Sabe-se que a utilização de pesos não binários nos termos de índices pode acarretar uma substancial melhoria na performance da recuperação de informação.

2.1.2 Modelo Vetorial

Segundo Salton (1971), o modelo vetorial utiliza-se de pesos não binários para o processo de recuperação de informação, propondo assim uma estrutura capaz de fazer uma combinação parcial entre a consulta e os documentos. Isto é realizado por uma atribuição de pesos não binários para os termos de índice de uma consulta e para os termos de índice dos documentos. Os pesos dos termos são utilizados para que se possa calcular o grau de similaridade entre cada documento armazenado no sistema e a consulta do usuário. Por classificação, a recuperação de documentos se dá em ordem decrescente ao grau de similaridade. O principal efeito resultante é uma resposta à consulta através de um conjunto de documentos ordenados, muito mais preciso que o conjunto de documentos recuperados utilizando-se o modelo booleano (Baeza-Yates & Ribeiro-Netto, 1999).

Um documento d_j e a consulta *q* de um usuário são representados por vetores *t-dimensionais*, como podemos ver na Figura 5 (Baeza-Yates & Ribeiro-Netto, 1999). Onde o cosseno do ângulo θ é o grau de similaridade entre a consulta d_j e o documento *q*.

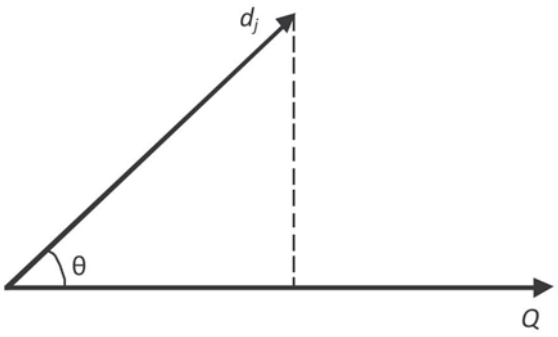

Figura 5 – O coseno de Θ é adotado como *sim(d_j,q)* (Baeza-Yates & Ribeiro-Netto, 1999)

De acordo com Baeza-Yates e Ribeiro-Netto (1999), a avaliação do grau de similaridade entre o documento d_j e a consulta *q* é proposto pelo modelo vetorial, sendo a correlação entre os vetores d_j e q. Esta correlação pode ser quantificada, pelo cosseno do ângulo entre estes dois vetores. Sendo:

$$sim(d_j) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times \|\vec{q}|} \tag{2.1}$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}} \tag{2.2}$$

Onde d_j representa o vetor do documento *d* e *q* representa uma consulta. O *ranking* dos documentos não é influenciado por *q* , uma vez que o fator é o mesmo para todos os documentos. O fator d_j fornece a normalização do documento no espaço.

Sabe-se que a partir de $w_{i,j} >= 0$ e $w_{i,q} >= 0$, a similaridade $sim(q, d_j)$ varia entre 0 e 1. Assim, o modelo vetorial organiza um *ranking* de documentos de acordo com o grau de similaridade entre a consulta e o documento, não se preocupando em tentar descobrir se o documento é ou não relevante para a consulta. Para melhorar a recuperação, pode-se determinar um limiar para o grau de similaridade, para que se recuperem somente os documentos acima de tal limiar, desprezando os documentos que não tenham alcançado o limiar estabelecido. Entretanto, para que se possa fazer um *ranking* dos documentos recuperados, deve-se explicitar antes como serão obtidos os pesos dos termos.

Salton e MacGill (1983), demonstram várias maneiras de se calcular o peso de um termo de índice. Neste trabalho serão tratados apenas princípios básicos que dão suporte às técnicas de *clustering*.

Tais técnicas são descritas como: tendo uma coleção C de objetos e uma pequena descrição de um conjunto A, tem-se como objetivo para um algoritmo de *clustering*, separar a coleção C

em dois conjuntos: (i) o algoritmo precisa determinar que conjunto é composto por objetos pertencentes ao conjunto A. É necessária uma pequena descrição do conjunto, pois não se possui uma informação completa para decidir precisamente quais os objetos fazem ou não fazem parte do conjunto A; (ii) o algoritmo precisa determinar quais são as características que melhor diferenciam os objetos do conjunto A dos objetos restantes da coleção C. O primeiro conjunto de características é fornecido para a quantificação de um *cluster* interno. Um algoritmo eficiente de *clustering* tenta balancear estes dois efeitos.

A similaridade de um *clustering* interno, no modelo vetorial é quantificada pela freqüência de um termo k_i dentro de um documento d_j.

Segundo Salton e MacGill (1983), a freqüência do termo utilizada é chamada de *tf* (do inglês *term frequency*) e provê uma medida da capacidade do termo de descrever o conteúdo de um documento (caracterização interna do documento). Além disso, a dissimilaridade interna do *cluster* é determinada pela freqüência inversa do documento (*idf* do inglês *inverse document frequency*) k_i entre os documentos de uma coleção. Utiliza-se o *idf* uma vez que termos que aparecem em muitos documentos não são bons para representar um documento de modo que se possa definir se o documento é relevante.

Definição: de acordo com Baeza-Yates e Ribeiro-Netto (1999), seja um número N o número total de documentos no sistema e seja n_i o número de documentos nos quais o termo de índice k_i esteja presente. Seja $freq_{i,j}$ a frequência do termo k_i no documento d_j (o número de vezes que o termo k_i é mencionado no texto no documento d_j). Então, a freqüência normalizada $f_{i,j}$ do termo k_i no documento d_j pode ser dada por:

$$f_{i,j} = \frac{freq_{i,j}}{max_l freq_{l,j}} \tag{2.3}$$

Onde o *max* é calculado sobre todos os termos que são mencionados no texto de um documento d_j. Se o termo k_i não aparece no documento d_j então $f_{i,j} = 0$. Adicionalmente, idf_i, a freqüência inversa do documento pelo k_i é dada por:

$$idf_i = \log \frac{N}{n_i} \tag{2.4}$$

Uma alternativa para se obter um peso com maior precisão para o termo é dada por Baeza-Yates e Ribeiro-Netto (1999) utilizando-se de:

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \tag{2.5}$$

Concluindo, as principais vantagens do modelo vetorial são: a utilização de peso para os termos melhora o desempenho da recuperação; a utilização da estratégia de comparação parcial permite a recuperação dos documentos que estão próximos à condição da consulta; sua fórmula de *ranking* utilizando o cosseno classifica os documentos de acordo com o seu grau de similaridade em relação a consulta. Teoricamente a desvantagem do modelo vetorial é o fato de assumir

que os termos de índice são mutuamente independentes. Entretanto, na prática, considerar-se um termo dependente pode ser uma desvantagem (Baeza-Yates & Ribeiro-Netto, 1999). Devido à dependência de muitos termos, sua aplicação indiscriminada para todos os documentos da coleção pode diminuir sua performance (Baeza-Yates & Ribeiro-Netto, 1999).

2.1.3 Modelo Probabilístico

O modelo probabilístico de recuperação de informação foi definido por Robertson e Spark Jones (1976). Mais tarde esse modelo viria a ser conhecido como modelo de recuperação de independência binária.

O modelo probabilístico visa recuperar informação utilizando uma estrutura probabilística (Baeza-Yates & Ribeiro-Netto, 1999). A recuperação de informação utilizando o modelo probabilístico, tem como idéia principal recuperar o documento relevante de forma exata e nada além disso. De posse da descrição exata do conjunto de resposta, a recuperação destes documentos torna-se uma tarefa trivial. Com isso, Baeza-Yates e Ribeiro-Netto (1999) definem que a ação de realizar uma consulta, nada mais é do que o ato de especificar as propriedades de um conjunto de respostas ideais. Infelizmente as propriedades não são conhecidas por nós, sendo conhecido apenas que os termos de índice são utilizados para caracterizar tais propriedades.

O usuário pode ajudar a identificar o conjunto de respostas ideais examinando o resultado da recuperação de documentos e decidindo quais documentos são relevantes e quais não são relevantes. O sistema então usa essas informações para refinar a descrição do conjunto ideal de respostas. Pela repetição desse processo espera-se alcançar um conjunto o mais perto possível do conjunto ideal de respostas.

O modelo probabilístico é baseado na seguinte suposição fundamental (Baeza-Yates & Ribeiro-Netto, 1999):

Supõe-se que dada uma consulta q de um usuário e um documento d na coleção, o modelo probabilístico tenta estimar a probabilidade do usuário encontrar o documento d, ou algum outro documento relevante a sua busca. O modelo probabilístico assume, que a relevância da probabilidade tem relação direta com a representação do documento e com a consulta. Entretanto o modelo assume que estes documentos são um subconjunto de todos os documentos que o usuário deseja, para um conjunto de respostas que satisfaçam uma consulta q. Um conjunto ideal de respostas é etiquetado com R e este conjunto deve melhorar a probabilidade de relevância para o usuário. Documentos do conjunto R são determinados como sendo relevantes de acordo com a consulta. Documentos que não estão presentes no conjunto R são referidos como não-relevantes à consulta.

Tendo-se uma consulta q, o modelo probabilístico atribui a cada documento d a medida da similaridade com a consulta, a relação P (probabilística) mostra a relevância ou não de d(consulta) para q(documento). Com isso calculando a probabilidade de d ser relevante para

a consulta q. Examinando-se as probabilidades de relevância utilizando-se um *ranking* de documentos pode-se minimizar julgamentos errados (Fuhr, 1992), (S. E. Robertson & Porter, 1981).

Definição: Baeza-Yates e Ribeiro-Netto (1999) definem que para o modelo probabilístico, que os pesos dos termos de índices são binários, por exemplo, $w_{i,j} \in \{0,1\}$, $w_{i,q} \in \{0,1\}$. Uma consulta *q* é um subconjunto dos termos de índice. Desde que *R* seja um conjunto de documentos conhecidos (ou supostamente conhecidos) como relevantes, o vetor $\overline{R}$ seja o complemento de *R* (i.e., o conjunto de documentos não relevantes). $P(R|\vec{d_j})$ seja a probabilidade desse documento d_j ser relevante para a consulta *q* e $P(\overline{R}|\vec{d_j})$ seja a probabilidade de este documento d_j ser relevante para q. A similaridade $sim(d_j, q)$ entre o documento d_j e a consulta *q* é definida por:

$$sim(d_j, q) = \frac{P(R|\vec{d_j})}{P(\overline{R}|\vec{d_j})} \tag{2.6}$$

Usando as regras de Bayes temos:

$$sim(d_j, q) = \frac{P(\vec{d_j}|R) \times P(R)}{P(\vec{d_j}|R) \times P(\overline{R})} \tag{2.7}$$

$P(\vec{dj}|R)$ é a probabilidade para a seleção randômica de documentos d_j de um conjunto R de documentos relevantes. P(R), é a probabilidade do documento selecionado randomicamente ser relevante dentro da coleção. $P(R) \cup P(\overline{R})$ é calculado para todos os documentos na coleção. Baeza-Yates e Ribeiro-Netto (1999) descreve:

$$sim(d_j, q) \frac{P(\vec{d_j}|R)}{P(\vec{d_j}|\overline{R})} \tag{2.8}$$

Assumindo a independência dos termos de índice, temos:

$$sim(d_j, q) = \frac{(\prod_{gi(\vec{d_j})=1} P(k_i|R)) \times (\prod_{gi(\vec{d_j})=1} P(\overline{k_i}|R))}{(\prod_{gi(\vec{d_j})=1} P(k_i|\vec{R})) \times \prod_{gi(\vec{d_j})=1} P(\overline{k_i}|\overline{R}))} \tag{2.9}$$

$P(k_i|R)$ é a probabilidade do termo de índice k_i estar presente na seleção randômica dos documentos do conjunto R. $P(\overline{k_i}|R)$ é a probabilidade do termo de índice k_i não estar presente na seleção randômica de um conjunto R. As probabilidades associadas ao conjunto $\overline{R}$ têm significados análogos para a descrição. Examinando os algoritmos, e tendo $P(k_i|R) + P(\overline{k_i}|R) = 1$ e ignorando fatores o quais são constantes para todos documentos no contexto das mesmas consultas, nós podemos finalmente escrever (Baeza-Yates & Ribeiro-Netto, 1999):

$$sim(d_j, q) = \sum_{i=1}^{t} W_{i,q} \times w_{i,j} \times |(log \frac{P(k_i \| R)}{1 - P(k_i|R)} + log \frac{1 - P(k_i \| \overline{R})}{P(k_i|\overline{R})}|) \tag{2.10}$$

Esta é uma expressão importante no modelo probabilístico para se calcular o *ranking*. É necessário fazer ou planejar um método para calcular inicialmente as probabilidades $P(k_i|R)$ e $P(k_i|\overline{R})$ uma vez que no início não se conhece o conjunto R. Existem muitas maneiras de se obter o cálculo das probabilidades, esta é apenas uma das maneiras existentes. A seguir serão discutidas duas maneiras de se calcular as probabilidades.

No início do processo (imediatamente depois da especificação da consulta), não há nenhum documento recuperado. Assim sendo, pode-se assumir duas simplificações que são: (i)$P(k_i|R)$ é constante para todos os termos de índice e k_i (tipicamente, igual para 0,5), e (ii) assume-se que a distribuição dos termos de índice entre os documentos não relevantes pode ser aproximada pela distribuição dos termos de índice entre todos os documentos da coleção.

As simplificações são:

$$P(k_i|R) = 0,5 \tag{2.11}$$

$$P(k_i|\overline{R}) = \frac{n_i}{N} \tag{2.12}$$

Onde, como já definido n_i é o número de documentos que contêm o termo de índice k_i e N é o número total de documentos da coleção. Dado esta suposição inicial, segundo Baeza-Yates e Ribeiro-Netto (1999) nós podemos recuperar documentos que contenham termos da consulta e prove um *ranking* probabilístico inicial para eles. Depois disto, este *ranking* inicial é melhorado como segue.

Seja V um subconjunto de documentos inicialmente recuperados e ordenados por um modelo probabilístico. Assim um subconjunto ao topo do *ranking* pode ser definido, por instâncias, constituido de r documentos ordenados onde r é um limiar definido previamente.

Seja V_i um subconjunto de V composto de documentos de V e que contém o termo de índice k_i.

Podemos simplificar da seguinte maneira: V e V_i são usados para identificar o número de elementos neste conjunto. Para melhorar o *ranking* probabilístico, (a) é necessário melhorar as suposições utilizando $P(k_i)$ e $P(K_i|R)$ para a distribuição dos termos de índice k_i entre todos os documentos recuperados, e (b) podemos através de $P(k_i|\overline{R})$ considerar que todos os documentos não recuperados são documentos não relevantes. Baeza-Yates e Ribeiro-Netto (1999) assumindo isso, determinam que:

$$P(k_i|R) = \frac{V_i}{V} \tag{2.13}$$

$$P(k_i|\overline{R}) = \frac{n_i - V_i}{N - V} \tag{2.14}$$

Estes processos podem ser repetidos recursivamente. Assim, é possível melhorar as suposições para as probabilidades $P(k_i|R)$ e $P(k_i|\overline{R})$ sem assistência humana ao assunto (contrari-

ando a idéia inicial)(Baeza-Yates & Ribeiro-Netto, 1999).

A última fórmula para $P(k_i|R)$ e $P(k_i|\overline{R})$ onde, frente a valores pequenos de V e V_i são tratados na prática como (V = 1 e V_i = 0).

Para burlar este problema Baeza-Yates e Ribeiro-Netto (1999) sujerem que um fator de ajuste seja adicionado a freqüência, através:

$$P(k_i|R) = \frac{V_i + 0,5}{V + 1} \tag{2.15}$$

$$P(k_i|\overline{R}) = \frac{n_i - V_i + 0,5}{N - V + 1} \tag{2.16}$$

A utilização da constante 0,5 nem sempre é eficiente. Uma alternativa para se achar o valor da constate ideal (ou o mais próximo do ideal), é fracionar n_i/N. São fatores de ajustamento:

$$P(k_i|R) = \frac{V_i + \frac{n_i}{N}}{V + 1} \tag{2.17}$$

$$P(k_i|R) = \frac{N_i - V_i + \frac{n_i}{N}}{N - V + 1} \tag{2.18}$$

Baeza-Yates e Ribeiro-Netto (1999) finalizam o estudo a respeito do modelo clássico probabilísitico, explicitando as vantagens e desvantagem desse modelo.

A principal vantagem do modelo probabilístico, é que os documentos são ordenados em ordem decrescente em sua probabilidade de ser relevante para com a consulta. A desvantagem inclui: (i) a necessidade de uma suposta separação inicial dos documentos em conjuntos de documentos relevantes e não relevantes; (ii) o fato do método não levar em consideração a freqüência em que o os termos de índice ocorrem dentro dos documentos (todos os pesos são binários); e (iii) a adoção da independência assumida para os termos de índice.

2.2 Métricas para avaliação de Sistemas de Recuperação de Informação

Nesta seção abordaremos as métricas de avaliação de SRIs mais utilizadas.

Segundo Manning et al. (2008) para se mensurar a efetividade de um sistema de recuperação de informação é necessário uma coleção de testes que contenham três características:

1. Uma coleção de documentos;

2. Informações que se pretende encontrar expressáveis em forma de consultas;

3. Um conjunto de julgamentos de relevâncias, padronizadas em uma avaliação binária, relevante (1) ou não relevante (0) para cada par consulta-documento.

A proposta padrão para avaliação de SRIs envolve a noção de documentos relevantes e documentos não relevantes (Manning et al., 2008). De acordo com as necessidades dos usuários, um documento em uma coleção de teste recebe uma classificação binária onde se diz que o documento é relevante ou não relevante. A coleção de documentos de teste e a suíte de informações necessárias para o usuário devem possuir um tamanho razoável. A relevância é estimada para uma informação necessária, não uma consulta. Podemos exemplificar do seguinte modo:

"A informação de que a união da prática de exercícios físicos e o hábito de uma alimentação saudável são eficazes para uma vida mais longa."

Esta informação pode ser traduzida para uma consulta assim:

exercício **AND** físico **AND** alimentação **AND** saudável **AND** vida **AND** longa

Um documento é relevante se a consulta acima for dirigida à informação necessária, e não porque isto simplesmente contém todas as palavras na consulta. Esta distinção é muitas vezes confundida na prática, porque as informações necessárias não são claras. Se um usuário digitar em um mecanismo de busca *web* a palavra *Puma*, ele pode estar tentando encontrar um exemplar do felino *Puma*. Ou estar procurando informações sobre a marca de roupas esportivas com o mesmo nome. Ao se utilizar somente uma palavra à uma consulta, torna-se muito difícil para o sistema saber qual a informação que o usuário realmente necessita. Entretanto é certo que o usuário sabe qual a informação ele necessita, e ele pode julgar o resultado retornado com base na sua relevância para isso. Para avaliar o sistema, é necessário uma expressão clara de uma informação que o usuário necessita, o qual pode ser utilizado para o julgamento dos documentos retornados como relevantes e não relevantes. Para a definição da relevância ou não do documento retornado é utilizado um valor binário, 1 para relevante e 0 para não relevante. Isto é feito para a simplificação da avaliação da relevância do documento (Manning et al., 2008).

Muitos sistemas possuem vários pesos (conhecidos como parâmetros) que pode ser ajustado para afinar a performance do sistema.

Segundo Manning et al. (2008) é um erro relatar os resultados em uma coleção de teste que obteve seus parâmetros afinados para maximizar sua performance em uma coleção. Tal afinamento aumenta a expectativa da performance do sistema, porque os parâmetros serão maximizados para um conjunto de consultas em particular. Neste caso o procedimento correto é ter um ou mais coleções de teste, e afinar os parâmetros durante o desenvolvimento da coleção de teste. O responsável pela coleção de teste então executa o sistema com aqueles parâmetros na coleção de teste e relata os resultados para tal coleção como uma estimativa parcial de performance.

Para a avaliação de SRIs além das métricas apresentadas anteriormente, faz-se necessária a utilização de coleções de textos para testes. A seguir apresentaremos algumas coleções conhecidas da literatura e que são ou foram utilizadas para a avaliação do desempenho de SRI.

A Coleção *Cranfield* foi a primeira coleção de teste que permitiu mensurar quantitativamente a eficiência da recuperação de informações. A coleção *Cranfield*[1] começou ser coletada no início dos anos de 1950, contendo 1.398 resumos de notícias de jornal sobre aerodinâmica, um conjunto de 225 consultas, e um julgamento exaustivo de relevância para todas os pares (consulta-documento).

Text Retrieval Conference (TREC) foi desenvolvido pelo Instituto Nacional de Padrões e Tecnologia (NIST)[2] realizou inúmeros testes para avaliação de RI desde 1992, utilizando muitas trilhas com diferentes coleções de teste. As coleções de testes renderam durantes esses anos 6 CDs contendo 1,89 milhões de documentos. Os documentos em sua maioria mas não todos oriundos de artigos de notícias de jornais. Os documentos estão distribuídos em 450 temas distintos. Atualmente as TRECs foram descontinuadas, sendo sua posição tomada pelo CLEF. O Forum de avaliação de multilíngües (CLEF)[3], avalia a recuperação de informações em línguas européias em vários idiomas.

Recentemente o NIST tem realizado avaliações utilizando coleções de documentos em um tamanho muito maior do que o que era utilizado nas TRECs. O NIST incluiu em suas avaliações a coleção de páginas *web* denominada GOV2[4]. Esta coleção conta com 25 milhões de páginas *web*. A coleção GOV2 foi desenvolvida para avaliar SRIs de grandes empresas de busca *web*.

O projeto NTCIR[5] foi desenvolvido com várias coleções de teste de mesmo tamanho para a coleção TREC. Seu desenvolvimento foi focado em idiomas da Ásia Oriental e tradução de informações recuperadas, onde consultas são realizadas em um idioma em uma coleção que contenha um ou mais documentos de outros idiomas.

A agência de notícias Reuters[6] possui atualmente duas coleções desenvolvidas, (i) Reuters-21578 e (ii) Reuters-RCV1. A coleção Reuters-21578 é muito utilizada pra a classificação de textos e ela é constituída de 21.578 artigos de notícias. Reuters-RCV1, consiste de 806.791 documentos.

A Coleção *Newsgroups*[7] é muito utilizada para a classificação de textos. Esta coleção possui 1.000 artigos de cada um dos vinte grupos de notícias da *Usenet* (cada nome de um grupo de notícias sendo considerado uma categoria). Após a remoção dos artigos duplicados a coleção conta com 18.941 artigos.

[1]Para mais informações acesse https://dspace.lib.cranfield.ac.uk/
[2]Para mais informações acesse http://trec.nist.gov/
[3]Para mais informações acesse http://www.clef-campaign.org/
[4]Para mais informações acesse http://ir.dcs.gla.ac.uk
[5]Para mais informações acesse http://research.nii.ac.jp/ntcir/data/data-en.html
[6]Para mais informações acesse http://www.reuters.com/
[7]Para mais informações acesse http://people.csail.mit.edu/jrennie/20Newsgroups/

2.2.1 Avaliação de conjuntos de documentos não ordenados

As duas medidas mais freqüentes para se mensurar a eficácia de um SRI é o cálculo da precisão e abrangência (Baeza-Yates & Ribeiro-Netto, 1999). O cálculo dessas métricas é realizado sobre o resultado da consulta realizada.

Precisão (Pr) é a fração dos documentos recuperados que são relevantes, e é expressada pela seguinte fórmula:

$$Pr = \frac{DocumentosRelevantesRecuperados}{DocumentosRecuperados} \tag{2.19}$$

Abrangêngia (Ab) é a fração dos documentos relevantes que são recuperados, e é expressada pela seguinte fórmula:

$$Ab = \frac{DocumentosRelevantesRecuperados}{DocumentosRelevantes} \tag{2.20}$$

Precisão e Abrangência pode ser feita examinando a Tabela 1 (Avancini et al., 2006):

Tabela 1 – Tabela de contigência (Avancini et al., 2006)

	relevantes	não relevantes
recuperados	verdadeiro positivo (tp)	falso positivo (fp)
não recuperados	falso negativo (fn)	verdadeiro negativo (tn)

Podemos expressar o conteúdo da Tabela 1 utilizando as seguintes fórmulas:

$$P = tp/(tp + fp) \tag{2.21}$$

$$R = tp/(tp + fn) \tag{2.22}$$

Outra medida muito utilizada é chamada Medida F (do inglês *F-measure*)(Baeza-Yates & Ribeiro-Netto, 1999), que utiliza o produto entre Precisão e Abrangência, calculando assim a média harmônica entre Precisão e Abrangência. A Medida F pode ser empregada segundo a seguinte fórmula:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = (\frac{\beta^2 + 1)}{\beta^2 P + R} \tag{2.23}$$

Onde:

$$\beta^2 = \frac{1-\alpha}{\alpha} \tag{2.24}$$

Onde $\alpha \in$ [0,1] e assim $\beta^2 \in$ [0,$\propto$]. A Medida F balanceada utiliza igualmente Precisão e Abrangência, utilizando α=$\frac{1}{2}$ e β=1. A Medida F balanceada é comumente expressado como F_1, que é uma abreviação de $F_{\beta=1}$. Quando usado β=1, podemos apresentar a seguinte fórmula:

$$F_{\beta=1} = \frac{2PR}{P+R} \tag{2.25}$$

No entanto, mesmo utilizando um fator de ponderação esta não é a única opção. Valores de β < 1 dão ênfase à precisão, enquanto valores de β > 1 da ênfase à abrangência. Por exemplo, um valor de β = 3 ouβ = 5 pode ser utilizado se a abrangência é o que se busca.

Precisão, Abrangência e Medida F são inerentemente medidas entre 0 e 1, entretanto elas também são comumente escritas na forma de porcentagem, em uma escala entre 0 e 100.

2.2.2 Avaliação de conjuntos de documentos ordenados

Precisão, abrangência e medida-F são medidas baseadas em conjuntos. Eles são calculados usando um conjunto de documentos não ordenados. Precisamos estender estas medidas (ou definir novas medidas) se formos avaliar os resultados da recuperação ordenados presentes nos mecanismos de buscas atuais. No contexto de recuperação ordenada, o conjunto apropriado de documentos recuperados é naturalmente dado pelos *k* documentos recuperados ao topo da lista de documentos.

A curva Precisão-Abrangência é estudada da seguinte forma: se o (k+1)^{esimo} documento recuperado é considerado não relevante, então a abrangência é a mesma para *k* documentos ao topo da lista de documentos recuperados, mas a precisão cai. Se for relevante, então ambos, precisão e abrangência aumentam, e a curva sobe para a direita. Para amenizar esta característica da curva, é utilizada a precisão interpolada (do inglês *Interpolated Precision*)(Manning et al., 2008). A precisão interpolada (P$_{interp}$) em uma certo nível de abrangência *r* é definido como a mais alta precisão encontrada por qualquer nível de abrangência *r'*>= *r*:

$$P_{interp}(r) = Max_{r'>=r} P_{interp}(r') \tag{2.26}$$

Outra medida tradicional, utilizada na avaliação dos SRIs pelas TRECs é a Precisão Média Interpolada dos Onze-Pontos (do inglês Eleven-Point Intrepolated Average Precision) (Manning et al., 2008). Para cada informação necessária, a precisão interpolada é medida em 11 níveis de abrangência de 0,0; 0,1; 0,2; ...; 1,0.

Para cada nível de abrangência, calculamos a média aritmética da precisão interpolada para cada nível de abrangência.

Atualmente outras medidas são mais comuns. Um exemplo disso é a Precisão Média (do inglês *mean average precision* (MAP)) (Buckley & Voorhees, 2004). Entre as medidas de avaliação, MAP mostra-se especialmente boa em discriminação e estabilidade. Para cada informação buscada, a precisão média é a média do valor da precisão obtido para o conjunto dos *k* documentos ao topo da lista de documentos recuperados. Seja, o conjunto dos documentos relevantes $\{d_j,...,d_{mj}\}$ para uma informação necessária q_j e o conjunto de resultados recuperados

ordenados R_{jk} o maior resultado d_k, então:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Pr(R_{jk}) \tag{2.27}$$

Quando um documento relevante não é recuperado, o valor da precisão da equação acima assume o valor 0. Para uma única informação buscada, a precisão média aproxima-se da área sob a curva interpolada precisão-abrangência para o conjunto de consultas.

Utilizando MAP, os níveis de abrangências não são escolhidos, e não há interpolação. O valor MAP para a coleção de avaliação é a medida aritmética dos valores da média da precisão para cada informação buscada. A pontuação calculada para a MAP normalmente varia à cada informação buscada dentro de um mesmo sistema. Isto significa que um conjunto de necessidades de informação teste deve ser amplo e diversificado o suficiente para ser representativo do sistema de eficácia em diferentes consultas.

As medidas vistas até o momento, excetuando-se a medida MAP, têm seus fatores de precisão calculados para todos os níveis de abrangência. Entretanto para aplicações como buscas na *web* esta abordagem não é pertinente para o usuário (Manning et al., 2008). Para esses usuários o que interesse na realidade são quantos bons resultados são disponibilizados para eles na primeira página ou no máximo nas três primeiras páginas. Isto reduz a mensuração da precisão poucos níveis de resultados recuperados fixados entre 10 e 30 documentos. Esta medida é chamada de Precisão para *K* (do inglês *Precision at k*), onde *k* é o número de documentos. A vantagem dessa medida é a de não ser necessário saber previamente o tamanho do conjunto do corpus. A desvantagem dessa medida é que ela é menos estável que as demais medidas, pois, o número total de documentos relevantes para uma consulta influencia a precisão para *K* documentos.

Uma alternativa para o problema encontrado na medida Precisão para *K*, é a media R-Precisão (do inglês *R-Precision*). O conjunto de documentos relevantes pode ser incompleto, tais quando a relevância é formada pela criação do julgamento de relevância para os melhores *k* resultados de um sistema particular em um conjunto de experimentos. R-Precisão ajusta para o tamanho do conjunto de documentos relevantes. Se os documentos são relevantes para uma consulta, examinamos os melhores resultados relevantes de um sistema, e descobrir que *r* são relevantes, então por definição não somente é a precisão (e portanto R-Precisão) $r/|REL|$, mas a abrangência deste conjunto de resultados é também $r/|REL|$. Assim R-Precisão é idêntico a medida Ponto de Equilíbrio (do inglês Break-even Point). A medida Ponto de Equilíbrio é outra medida utilizada, definida em termos de relacionamentos explorados. Como a Precisão para *K*, R-Precisão descreve somente um ponto na curva Precisão-Abrangência, especialmente que tenta resumir a eficácia em toda a curva, uma vez que o usuário deve estar interessado em um ponto de equilíbrio e não deve querer saber o melhor ponto da curva (o ponto máximo utilizando a Medida-F) ou a recuperação do nível de interesse de uma aplicação específica (Precisão para *K*).

Outras duas medidas aplicada rotineiramente quando utilizado aprendizado de máquina para classificação de documentos é o Ganho Cumulativo (do inglês Cumulative Gain) e a medida Ganho Cumulativo Descontado Normalizado (do inglês *Normalized Discounted Cumulative Gain* (NDCG)). O NDCG é utilizado para situações onde não se utiliza a noção binária para a relevância de documentos. Como a Precisão para *K*, esta medida é avaliada sobre qualquer *K* ao topo do resultado. Sendo *R(j,d)* a pontuação de relevância dado pelos avaliadores para o documento *d* para a consulta *j*. Entao:

$$NDCG(Q,K) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_K \sum_{m=1}^{K} \frac{2^{R(j,m)}-1}{log(1+m)'} \tag{2.28}$$

Onde Z_K e um fator de normalização calculado. Este fator tem o intuito de torná-lo um perfeito ranking de NDCG sendo este 1 para *K*. Para consultas para cada *K'* < *K* documentos são recuperados e o último somatório é realizado até *K'*.

2.2.3 Avaliação de Relevância

Para a avaliação correta de um SRI, a informação que se busca encontrar tem que ser relevante para aos documentos existentes na coleção de documentos de teste, e deve ser apropriado para o sistema. As informações buscadas são melhores designadas por um especialista do domínio. Utilizando combinações aleatórias de termos de consultas como uma informação a ser recuperada geralmente não é uma boa idéia porque tipicamente estes termos não irão se parecer com a atual distribuição das informações buscadas.

Dadas as informações buscadas, é necessário coletar a avaliação da sua relevância. Este processo é demorado e dispendioso quando executado por seres humanos. Para coleções pequenas como a Coleção *Cranfield*, julgamento exaustivos de cada consulta e cada documento é realizada. Já para coleções maiores como são as coleções atuais, é usual o julgamento da avaliação da relevância ser realizado somente para um subconjunto dos documentos para cada consulta. Este método é denominado de *Pooling* (Buckley & Voorhees, 2004), onde a avaliação da relevância é realizada sobre um subconjunto da coleção que é formada por *K* documentos retornados melhor ranqueados.

Um humano pode não ser confiável para reportar um julgamento correto de relevância de um documento para uma consulta. Particularmente e seus julgamentos de relevância são totalmente particulares e variáveis. Mas isto não é problema para ser resolvido, na analise final, o sucesso de um SRI depende em quão bom esta avaliação satisfaz a necessidade destas particularidades humanas, uma informação buscada por vez.

Todavia, isto é interessante para considerar e medir o quanto há de entendimento entre julgadores e as informações julgadas relevantes. Na ciência social, uma medida comum de entendimento entre julgadores é a medida chamada Estatística Kappa (do inglês *Kappa Statistic*)

(Manning et al., 2008). Esta medida é designada para o julgamento absoluto e correto um simples acordo medido para avaliar a mudança de entendimento entre os julgamentos.

$$kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.29}$$

Onde P(A) é a proporção das vezes julgadas ajustadas, e P(E) é a proporção de vezes que ele espera concordar por mudança.

Tabela 2 – Cálculo da Estatística *Kappa* para o julgamento da relevância dos documentos (Manning et al., 2008)

		julgamento 2 relevância		
		sim	não	total
julgamento 1 relevância	sim	300	20	320
	não	10	70	80
	total	310	90	400

Na Tabela 2 podemos observar o entendimento entre os julgamentos 1 e 2:

P(A) = (300 + 70)/400 = 370/400 = 0,925

Limites conciliados

P(não relevantes) = (80 + 90)/(400 + 400) = 170/800 = 0,2125

P(relevantes) = (320 + 310)/(400 + 400) = 630/800 = 0,7878

A probabilidade de dois juízes concordarem pela mudança

P(E)[8] = P(não relevante) + P(relevante) = 0,2125 + 0,7878 = 0,665

Estatística *Kappa*

K = (P(A) - P(E))/(1 - P(E)) = (0,925 - 0,665)/(1 - 0,665) = 0,776

A avaliação do julgamento de relevância por parte dos julgadores humanos é utilizada em muitas coleções, como por exemplo, as TRECs e na coleções médicas. Utilizando para isso as regras mostradas na Tabela 2.

2.3 Considerações sobre o capítulo

Neste capítulo apresentamos uma revisão da literatura sobre Recuperação de Informação. Abordamos para este trabalho os modelos clássicos para o processo de recuperação de informa-

[8] A estatística limítrofe é calculada somando as linhas ou colunas.

ção. Os modelos clássicos de RI são: (i) modelo Booleano (Seção 2.1.1), (ii) modelo Vetorial (Seção 2.1.2) e (iii) modelo Probabilístico (Seção 2.1.3). Esta revisão foi muito importante para o melhor entendimento do funcionamento do processo de recuperação de informação e assim poder discernir melhor a respeito dos sistemas de recuperação de informação.

Neste capítulo também abordamos as principais métricas para mensurar o desempenho da recuperação de informação. Apresentamos também algumas coleções de documentos que são utilizadas com freqüência para a avaliação do processo de RI. Na Subseção 2.2.1 apresentamos as métricas de avaliação para o processo de recuperação de documentos não ordenados. Na Seção 2.2.2 apresentamos as métricas utilizadas para a avaliação de conjuntos o resultado do processo de recuperação de documentos ordenados. Este estudo foi muito importante para o processo de avaliação dos experimentos realizados nesta dissertação. Com este estudo podemos formular melhor a avaliação dos resultados obtidos além de conhecer outras métricas que são habitualmente utilizadas para avaliação da RI.

No próximo Capítulo (Capítulo 3) apresentaremos uma revisão da literatura sobre Expansão de Consultas (Seção 3.1) e Realimentação de Relevantes (Seção 3.2). Ao abordarmos as técnicas de EC apresentamos duas proposta de análise de documentos, Análise Automática Local e Análise Automática Global, seções 3.1.1 e 3.1.2 respectivamente. Ao apresentarmos a técnica de Realimentação de Relevantes, abordaremos sua aplicação em conjunto com os modelos clássicos de RI, Modelo Espaço Vetorial (Seção 3.2.1), Modelo Booleano (Seção 3.2.2) e Modelo Probabilístico (3.2.3). Na Seção 3.2.4 apresentamos o método de Pseudo Realimentação de Relevantes, este método é uma variação da Realimentação de Relevantes e visa automatizar a escolha das informações que serão adicionadas às consultas originais.

3 Expansão de Consultas e Realimentação de Relevantes

3.1 Expansão de Consultas

Na expansão de consulta (EC) , os usuários contribuem adicionando na consulta inicial palavras ou frases (Manning et al., 2008). Alguns mecanismos de buscas sugerem consultas relacionadas às respostas da consulta original (mecanismos de buscas *web*), cabendo ao usuário utilizar umas destas sugestões para expandir a consulta original. A Figura 6 mostra um exemplo de consulta realizada no mecanismo de busca web *Yahoo!*[1]. Na consulta realizada foi utilizado a palavra "*car*", e partindo dessa consulta o mecanismo pode sugerir novas palavras para a expansão da consulta.

A questão central nesta forma de expansão de consulta é como gerar alternativas ou consultas expandidas para o usuário. As formas mais comuns de expansão de consulta são a análise automática global e a análise automática local.

3.1.1 Análise Automática Local

Na estratégia de análise local, os documentos recuperados para uma certa consulta q são examinados no tempo da consulta para determinar os termos que serão utilizados na expansão da consulta (Baeza-Yates & Ribeiro-Netto, 1999). O processo é similar a pseudo realimentação de relevantes, que apresentaremos na Subseção 3.2.4, ou seja, não utiliza a iteração do usuário para a expansão da consulta.

Xu e Croft (1996) apresentam dois tipos de análise automática local, (i) realimentação local e (ii) análise do contexto local. Realimentação local (Attar & Fraenkel, 1977), utiliza os n documentos melhores ranqueados dos documentos recuperados pela consulta original para construir um *thesaurus* de forma automática. A informação dos documentos melhor classificados recuperados pela consulta inicial é utilizado para re-estimar a probabilidade de ocorrência no conjunto de documentos relevantes à consulta original (Croft et al., 1995), ou seja, não são adicionados novos termos e sim os pesos dos termos são modificados. A utilização desta proposta obteve bons resultados quando aplicados em coleções de documentos de testes pequenas (Xu & Croft, 1996). Uma versão simples da técnica de realimentação local onde palavras comuns dos documentos melhores ranqueados são adicionadas à consulta original. A eficiência desta versão

[1] Disponível em http://br.yahoo.com/

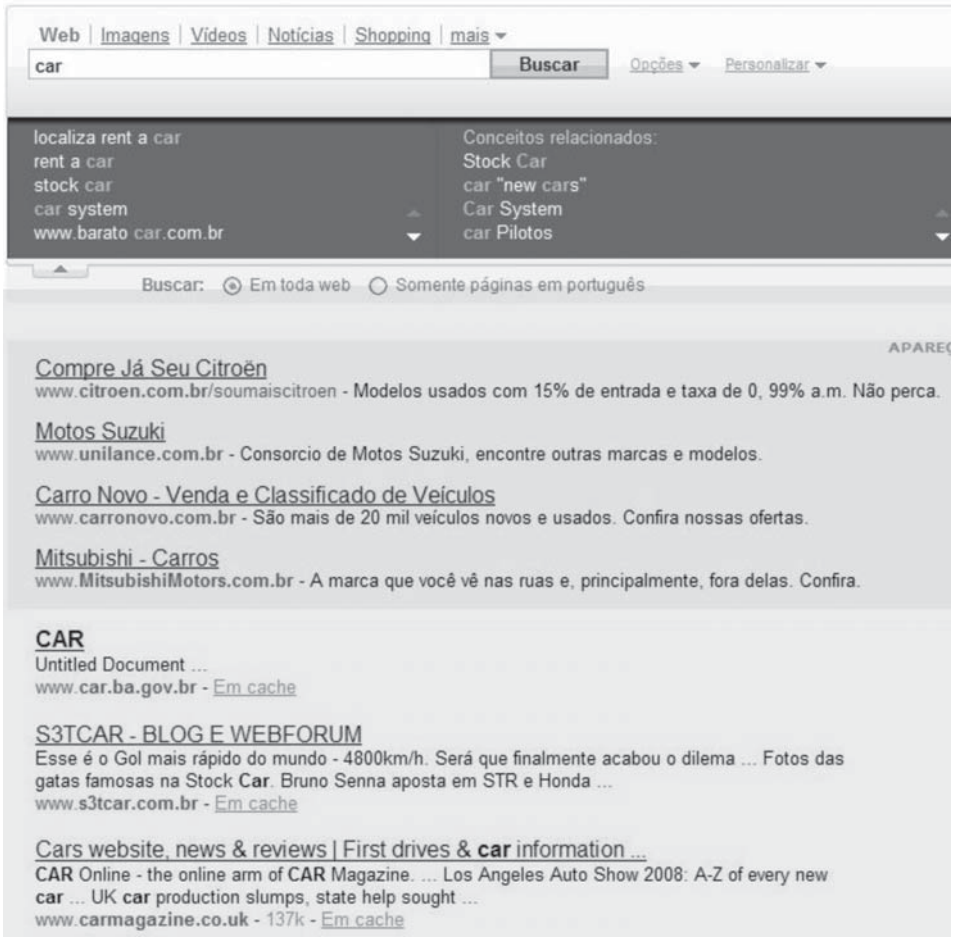

Figura 6 – Resultado da consulta para a palavra *car*

é altamente influenciada pela proporção de documentos relevantes no topo do ranque (Xu & Croft, 1996).

Uma vantagem da realimentação local é que ela pode ser relativamente eficiente para realizar a expansão baseada nos documentos no topo do ranque dos documentos recuperados pela consulta original (Xu & Croft, 1996). Isto pode ser levemente mais lento em tempo de execução, entretanto não é necessário a construção de um *thesaurus*. A realimentação local requer um busca e acesso da informação no documento extra. Se a informação do documento é armazenada somente para este propósito, então isto deverá ser contado como um espaço extra em disco para a técnica, entretanto, isto pode ser significativamente menor que uma base de dados de conceitos (Xu & Croft, 1996). A desvantagem da realimentação local é que ainda não tem muitas alternativas de se trabalhar quando as consultas recuperam poucos documentos relevantes.

Análise do contexto local é uma técnica que combina características tanto da análise global quanto da realimentação local (Xu & Croft, 1996). A análise do contexto local utiliza grupos de substantivos como conceitos e estes conceitos são selecionados baseados na co-ocorrência com os termos da consulta. Os conceitos são escolhidos dos documentos melhor ranqueados entre

os documentos recuperados pela consulta original (processo similar a realimentação local), mas as melhores passagens são utilizadas ao invés da totalidade dos documentos.

Conceitos (substantivos) nas *n* melhores passagens são ranqueados de acordo com a seguinte equação (Xu & Croft, 1996):

$$bel(Q,c) = \phi_{t_i \in Q}(\delta + \log(af(c,t_i))idf_c / \log(n))^{idf_i} \tag{3.1}$$

Onde:

- *af(c,t_i)* = $\sum_{j=1}^{j=n} \text{ft}_{ij}\ \text{fc}_j$;
- idf_i = max(1,0,log10(N/N_i)/5.0);
- idf_c = max(1,0,log10(N/N_c)/5.0).
- *c* é um conceito;
- ft_{ij} é o número de ocorrências de t_i em p_j;
- fc_j é o número de ocorrências de *c* em p_j
- *N* é o número de passagens na coleção;
- N_i é o número de passagens contendo t_i;
- N_c é o número de passagens contendo *c*;
- δ é 0,1 para o valor de "*bel*" igual a 0.

A fórmula acima é uma variação da medida *tf idf* (Salton & Buckley, 1988). Na fórmula, *af* recompensa os conceitos onde os termos da consulta co-ocorrem freqüentemente, o idf_c penaliza conceitos que ocorrem muito na coleção, o idf_i enfatiza termos da consulta com baixa freqüência. A multiplicação da ênfase a co-ocorrência com todos os temos da consulta.

Análise do contexto local possui muitas vantagens, entre eles estão a sua praticidade computacional (Xu & Croft, 1996). Para cada coleção é necessário somente uma única passagem para coletar a freqüência dos termos e dos substantivos. A principal desvantagem da análise do contexto local é que este pode requerer um tempo grande para a expansão das consultas (Jing & Croft, 1994).

3.1.2 Análise Automática Global

A idéia básica na análise global é que o contexto global de um conceito pode ser utilizado para determinar a similaridade entre conceitos (Xu & Croft, 1996). Contexto como conceitos

podem ser definidos de várias formas. Em uma definição simplista, todas as palavras são conceitos (excetuando-se as palavras definidas como *stopwords*) e que o contexto para uma palavra é toda palavra que co-ocorre em um documento com tal palavra (Qiu & Frei, 1993). O principal diferencial da análise global, é que ela é usada somente para expansão de consulta, e não substitui a representação original do documento baseada nas palavras (Xu & Croft, 1996). Crouch e Yang (1992) apresentam uma proposta com a utilização de agrupamentos para determinar a análise do contexto para o documento.

Uma das primeiras técnicas que apresentou resultados consistentes e efetivos foi a análise de global aplicada ao sistema INQUERY (Callan et al., 1995).A técnica utilizada no sistema INQUERY, determina que conceito como sendo um grupo de substantivos (um, dois ou três substântivos adjacentes), e o contexto é definido como uma coleção de tamanho determinado (uma janela) em torno dos conceitos (Qiu & Frei, 1993). Uma janela para ser efetiva possui entre uma e três sentenças. Um caminho para a visualização da técnica, entretanto de difícil implementação, é o caminho de considerar cada conceito (grupo de substantivo) para ser associado com um pseudo documento. O conteúdo do pseudo documento, são as palavras que ocorrem em cada janela a tal conceito no corpus. Por exemplo, o conceito *airline pilot* pode ter as palavras *pay*, *strike*, *safety*, *air*, *traffic* e *FAA* ocorrendo freqüentemente no pseudo documento correspondente, de acordo com o corpus analisado. O banco de dados do INQUERY é construído a partir dos pseudo documentos, criando um bando de dados de conceitos.

A principal vantagem da proposta expansão de consulta com análise global, é que ela é relativamente robusta no que tendem o desempenho médio das consultas para este tipo de expansão (Xu & Croft, 1996). A desvantagem da proposta expansão de consulta com análise global é que a proposta pode ser muito onerosa em se tratando de espaço de disco e tempo computacional para a análise do contexto global e construir uma base de dados pesquisável, e consultas individuais podem ser significativamente degradadas para a expansão (Xu & Croft, 1996).

3.2 Realimentação de Relevantes

A Realimentação de Relevantes (RR) (*Relevance Feedback* - RF) é um processo automático para a modificação da consulta inicial em um SRI com base no julgamento da relevância dos documentos recuperados anteriormente (Salton, 1971). A idéia por trás da RR é envolver o usuário no processo de RI para melhorar o resultado final da recuperação. Em particular os usuários devem julgar a relevância dos documentos recuperados em um resultado preliminar (Manning et al., 2008). O processo de RR melhora a formulação da consulta escolhendo termos importantes de documentos considerados relevantes pelo usuário recuperados pela consulta original. Realimentação de Relevantes prevê a participação do usuário para identificar informações importantes à consulta original (Baeza-Yates & Ribeiro-Netto, 1999). Estas informações podem ser documentos inteiros ou partes dos mesmos. Do conjunto de documentos inteiros ou

partes destes julgados como relevantes pelo usuário é feita a extração dos termos ou expressões que incrementarão a consulta original, gerando assim uma nova consulta (White & Marchionini, 2006). O processo RR é um método interativo que pode ser repetido quantas vezes forem necessários, até o momento em que o usuário estiver satisfeito com o resultado da consulta (Orengo & Huyck, 2006). Os procedimentos básicos para RR são (Manning et al., 2008):

- O usuário realiza uma consulta;
- O sistema retorna um conjunto de documentos como resultado inicial;
- O usuário marca alguns dos documentos retornados como relevantes ou não relevantes
- O sistema calcula a melhor representação da informação pesquisada com base na realimentação do usuário;
- O sistema mostra ao usuário um conjunto revisado do resultado recuperado.

Na Figura 7, podemos observar o processo de RR para melhorar a consulta original.

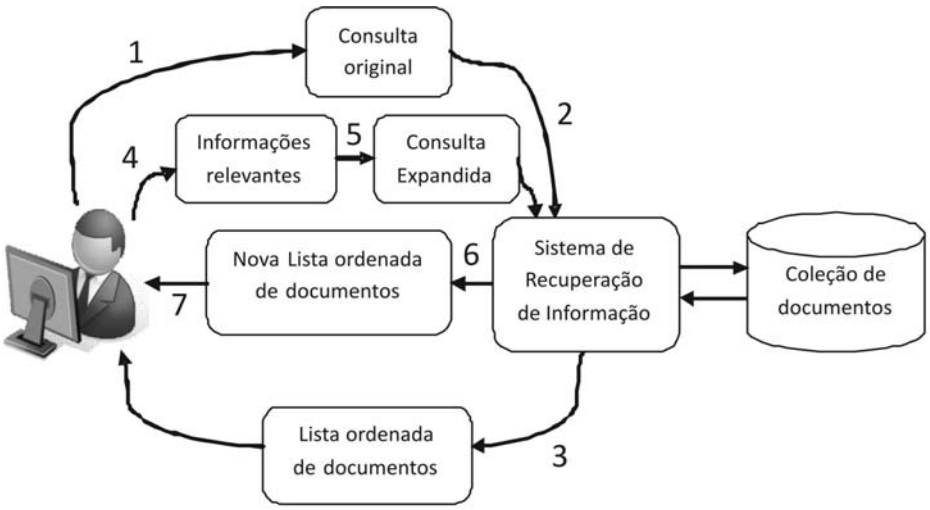

Figura 7 – Processo de Realimentação de Relevantes

As principais vantagens da RR são a sua simplicidade e bons resultados (Baeza-Yates & Ribeiro-Netto, 1999). Outras vantagens para a utilização do processo de realimentação de relevantes (Salton & Buckley, 1997):

- Auxilia os usuários através do processo de reformulação da consulta;
- A operação de busca é realizada em uma seqüência de pequenos passos para aproximar o resultado da consulta como o que o usuário necessita;
- Permite um processo controlado das alterações da consulta, visando destacar certos termos, como requer ambientes de busca em particular.

O conceito de realimentação de relevantes foi introduzido em meados da década de 1960 (Orengo & Huyck, 2006). A RR pode ser aplicada utilizando diversos modelos de RI. Neste trabalho abordaremos a utilização de RR nos modelos clássicos de RI.

3.2.1 Realimentação de Relevantes no Modelo Espaço Vetorial

O processo de realimentação de relevantes orginalmente, foi desenvolvido para ser aplicado utilizando vetores de consultas. Estes vetores são definidos sem a presença de operadores Booleanos (Orengo & Huyck, 2006). O processo orginal de RR pode ser apresentado da seguinte forma:

$$Q_0 = (q_1, q_2, ..., q_n) \tag{3.2}$$

Onde q_1 representa o peso do termo 1 na consulta Q_0. O peso dos termos variam entre 0 a 1. O termo com peso 0, indica que ele não está presente na consulta, em contra ponto, o termo valorado com peso igual a 1, está fortemente presente na consulta.

O processo de realimentação de relevantes dará origem a uma nova versão da consulta original (Orengo & Huyck, 2006), que pode ser apresentada da seguinte forma:

$$Q'_0 = (q'_1, q'_2, ..., q'_n) \tag{3.3}$$

Onde q'_1 representa o peso referente ao termo 1 modificado. A adição de novos termos podem ser exercida modificando o peso de 0 para um valor maior que 0. Este processo de valoração do termos busca aproximar o vetor da consulta de documentos relevantes, ao mesmo tempo em que ele se deva se afastar dos documentos não relevantes para a consulta.

Harman (1992) apresentou uma revisão do processo de realimentação de relevantes. Nesta revisão do processo de RR, é apresentado três métodos para a utilização da realimentação de relevantes inserida no Modelo Espaço Vetorial. Estes três métodos basicamente unem os vetores dos documentos aos vetores da consulta inicial. Os três métodos apresentados são:

- *Ide Regular*

$$Q' = Q_0 + \sum_{k=1}^{n_1} R_k - \sum_{k=1}^{n_2} S_k \tag{3.4}$$

- *Ide dec-hi*

$$Q' = Q_0 + \sum_{k=1}^{n_1} R_k - S_1 \tag{3.5}$$

- *Rocchio*

$$Q' = Q_0 + \beta \sum_{k=1}^{n_1} \frac{R_k}{n_1} - \gamma \sum_{k=1}^{n_2} \frac{S_k}{n_2} \tag{3.6}$$

Onde:

- R_k é o vetor para o documento relevante *k*;
- S_k é o vetor para o documento não relevante *k*;
- n_i é o número de documentos relevantes;
- n_2 é o número de documentos não relevantes;
- β e γ são parâmetros que controlam a contribuição dos documentos relevantes e não relevantes.

Como exposto anteriormente, os três métodos tem como procedimentos básicos a união dos vetores dos documentos e os vetores da consulta original. Esta união dos vetores realiza uma nova definição dos pesos dos termos da consulta original de forma automática. Este processo soma o novo peso ao peso da ocorrência atual dos termos da consulta nos documentos relevantes. Em contrapartida a nova definição dos pesos, subtrai do peso dos termos da consulta original os pesos dos termos da consulta presentes nos documentos não relevantes. A consulta original tem seus termos expandidos pela adição automática dos termos presentes nos documentos relevantes e não relevantes e que não ainda não estavam presentes na na consulta original.

A expansão das consultas é realizada, utilizando somente pesos oriundos de documentos relevantes. Os termos dos documentos julgados não relevantes participa do processo de expansão modificando os pesos dos novos termos vindos de documentos relevantes (Harman, 1992).

O método *Ide dec-hi* utiliza somente documentos julgados não relevantes para a realimentação. São utilizados os documentos melhor ranqueados, descartando os documentos não relevantes recuperados e apresentados ao usuário (Harman, 1992).

O método *Rocchio*, utiliza é baseado na normalização do peso dos documentos, este esquema é utilizado tanto em documentos relevantes como para os documentos não relevantes (Harman, 1992).

Salton e Buckley (1997) apresentaram uma comparação dos três métodos em experimentos realizados em dois níveis de expansão de consulta em seis corpus diferentes. O melhor método para todo os seis corpus foi o *Ide dec-hi*, embora a diferença para os demais métodos seja pequena. Para o método *Rocchio*, o melhor resultado foi alcançado utilizando γ = 0,25 e β = 0,75, limitando com isso o efeito da realimentação negativa, sendo isto feito de forma automática pelo método *Ide dec-hi*.

3.2.2 Realimentação de Relevantes no Modelo Booleano

Ao compararmos a utilização do processo de realimentação de relevantes aplicado ao modelo Booleano com os demais modelos clássicos (Espaço Vetorial e Probabilístico), podemos dizer que o primeiro é utilizado em menor escala do que os demais. Segundo Orengo (2004),isto acontece porque na realimentação de relevantes a escolha dos termos é crucial e no modelo Booleano é necessário a escolha dos operadores para que se possa relacionar os termos.

O modelo Booleano de realimentação utiliza informações relevantes providas pelo usuário para calcular o peso para os termos nos documentos recuperados (Dillon et al., 1983). Estes pesos são utilizados para ordenadar os termos, buscando a construção de uma nova consulta booleana. Os valores utilizados pelos pesos se limitam ao intervalo entre -1 a 1. A nova consulta será constituída de duas partes, (i) uma parte com termos com pesos altos, considerados bons termos e (ii) outra parte com termos com pesos baixos, considerados termos ruins. Tanto os termos considerados bons e ruins, são divididos em dois grupos. Termos bons são divididos em: (i) termos do primeiro grupo (os melhores classificados) sendo, *(t_1 ou t_2 ou ...ou t_n)*; (ii) termos do segundo grupo em pares *(t_1 e t_2)*, cada termo é colocado em pares com qualquer outro termo no grupo. Termos ruins são dividos pela regra, onde os termos do pior grupo são utilizados da seguinte forma: (i) *não t_1 e não t_2 e ...e não t_n*; (ii) termos do segundo pior grupo em pares na forma (t_1 *e* t_2), para todos os pares.

O método de realimentação de relevantes utilizando o modelo Booleano é constituído segundo Salton e MacGill (1983) em duas etapas: (i) etapa de construção de boas cláusulas e (ii) etapa de construção da consulta Booleana modificada, utilizando algumas cláusulas previamente escolhidas. A utilização da cláusula (que pode ser um termo único ou um conjunto de termos conectados pelo operador "*e*") depende do seu peso de relevância e da transferência da freqüência esperada. A medida do peso de relevância é importante para saber se a clausula é útil para recuperar documentos relevantes. A realimentação da consulta consiste de um conjunto de cláusulas conectadas pelo operador lógico *ou* (Orengo, 2004).

3.2.3 Realimentação de Relevantes no Modelo Probabilístico

O modelo probabilístico é baseado na idéia apresentada por Robertson e Spark Jones (1976) da distribuição dos termos da consulta em relevantes e não relevantes. Esta distribuição é realizada, definindo-se os pesos dos termos, da pontuação dos documentos recuperados, e pela soma entre os pesos dos termos presentes nos documentos presentes na consulta (Harman, 1992). A definição do peso dos termos é realizada pela seguinte fórmula (Robertson & Spark Jones, 1976):

$$w_{ij} = \log_2 \frac{\frac{r}{R-r}}{\frac{n-r}{N-n-R-r}} \tag{3.7}$$

Onde:

- w_{ij} = o peso do termo *i* para a consulta *j*;
- N = o número de documentos na coleção;
- R = o número de documentos relevantes para a consulta *j*;
- n = o número de documentos que possuem o termo *i*;
- r = o número de documentos relevantes que possuem o termo *i*.

Jones (1997) apresenta um experimento similar a utilização da fórmula de pesagem de relevantes em uma situação operacional de realimentação de relevantes, na qual o usuário verifica somente alguns documentos relevantes em um conjunto inicial de documentos recuperados, e daqueles poucos documentos são somente disponíveis para o esquema de pesagem. O resultado desta nova pesagem com somente alguns documentos relevantes mostrou melhora significante no seu desempenho em comparação com a performance da definição de novos pesos utilizando somente a medida *IDF*(Salton & MacGill, 1983). Isto indica que o esquema de nova pesagem probabilística provê um método eficaz para realimentação de relevantes especialmente na nova pesagem dos termos (Harman, 1992).

A principal vantagem da utilização do modelo probabilístico em conjunto com a realimentação de relevantes, segundo Baeza-Yates e Ribeiro-Netto (1999), é que o processo de realimentação de relevantes é diretamente relacionado para a derivação de novos pesos para os termos da consulta. Suas desvantagens são: (ii) a definição dos pesos dos termos dos documentos não são realizados na iteração do processo de realimentação; (ii) pesos calculados em formulações de consultas anteriores são desprezadas; (iii) não é utilizada em expansão de consulta, somente termos presentes na consulta inicial são pesados novamente.

Para uma recuperação mais eficiente, na abordagem de RR utilizando o modelo Probabilístico, é utilizada a ordenação dos documentos em forma decrescente de acordo com a seguinte fórmula (Salton & Buckley, 1997):

$$\log \frac{Pr(x|rel)}{Pr(x|nonrel)} \tag{3.8}$$

Onde: *Pr(x|rel)* e *Pr(x|nonrel)* são a probabilidade da representação de um item relevante ou não no vetor *x*.

A definição dos termos é realizada independentemente da relevância dos documentos da coleção. Os pesos dos termos atribuidos aos documentos são definidos utilizando valores binários 0 e 1. Para o cálculo da similaridade entre a consulta e o documento, podemos utilizar a derivação da equação 3.8, aplicando-a à consulta e cada documento $D = (d_1, d_2,...,d_t)$, através de dois

parâmetros (p_i e u_i) que representam a probabilidade que o *i-ésimo* termo tenha um valor 1 e um documento relevante ou não (Salton & Buckley, 1997). Equação 3.8 derivada é apresentada da seguinte forma:

$$sim(Q, D) = \sum_{i-1}^{t} d_i \log \frac{p_i(1 - u_i)}{u_i(1 - p_i)} + constante \tag{3.9}$$

Onde:

$$\mathrm{p}_i = \mathrm{Pr}(\mathrm{x}_i = 1 \mid \text{relevante})$$
$$\mathrm{u}_i = \mathrm{Pr}(\mathrm{x}_i = 1 \mid \text{não relevante})$$

O cálculo da similaridade (a fórmula 3.9) entre a consulta e os documentos, não pode ser utilizada na prática sem o conhecimento prévio para todos os termos do documento dos valores de p_i e u_i. Segundo Salton e Buckley (1997) alguns métodos foram apresentados para o cálculo dos valores de p_i e u_i. Para a pesquisa inicial, quando ainda não se tem conhecimento da relevância das informações dos documentos, assume-se que o valor para p_i é constante e geralmente 0,5.

A Tabela 3 apresenta a ocorrência do termo *i* em um subconjunto de documentos relevantes e não relevantes, u_i pode ser definido o equivalente n_i/N, a proporção dos documentos na coleção que possui o termo *i*. Para a rodada inicial, a expressão 3.9 é então reduzida para (Salton & Buckley, 1997):

$$sim - inicial(D, Q) = \sum_{i=1}^{t} d_j \log \frac{N - n_i}{n_i} \tag{3.10}$$

No contexto da realimentação das consultas, os valores acumulados e relacionados à relevância dos itens recuperados são utilizados para avaliar a fórmula 3.9. A avaliação é realizada pela distribuição do termo nos itens relevantes recuperados anteriormente. Esta distribuição é a mesma para todo o conjunto de itens relevantes, sendo os itens não recuperados rotulados como não relevantes (Salton & Buckley, 1997). Aplicando os fatores presentes na Tabela 3 para a os documentos recuperados da coleção, temos que:

$$p_i = \frac{r_i}{R} \tag{3.11}$$

$$u_i = \frac{n_i - r_i}{N - R} \tag{3.12}$$

Salton e Buckley (1997) apresenta uma variação da fórmula 3.9, substituindo pi e u_i utilizando as expressões 3.11 e 3.12.

$$sim(Q, D) = \sum_{i=1}^{t} d_j \log(\frac{r_i}{R - r_i} / \frac{n_i - ri}{N - R - n_i - r_i}) \tag{3.13}$$

Tabela 3 – Ocorrência do termo *i* na coleção de documentos *N* (Salton & Buckley, 1997)

	Itens Relevantes	Itens Não Relevantes	Todos os Itens
d_i=1	r_i	n_i - r_i	n_i
d_i=0	R - r_i	N - R - n_i + r_i	N - n_i
Todos os Itens	R	N - R	N

Onde na fórmula 3.13 *R* representa o número total de itens relevantes recuperados, r_i é o número total de itens relevantes recuperados que possuem o termo *i*, e n_i é o número total de itens recuperados que possuem o termo *i*.

Salton e Buckley (1997), para alguns valores muito pequenos para *R* e r_i A fórmula 3.13, pode causar alguns problemas. Estes problemas freqüentemente acontecem na prática (exemplo: *R* = 1 e r_i = 0), por causa da expressão logarítmica é então reduzida à 0 (Salton & Buckley, 1997). Para amenizar este problema, muitas vezes um fator de ajuste (0,5) é adicionado na definição de p_i e u_i. Com isso as fórmulas 3.14 e 3.15 são utilizados em sistemas probabilísticos convencionais para a obtenção dos valores de p_i e u_i (Salton & Buckley, 1997).

$$p_i = \frac{r_i + 0,5}{R + 1} \tag{3.14}$$

$$u_i = \frac{n_i - r_i + 0,5}{N - R + 1} \tag{3.15}$$

Entretanto segundo Salton e Buckley (1997), o fator de ajuste nem sempre é satisfatório, para estes casos, utiliza-se como alternativa o calculo do valor de p_i e u_i tal que, n_i/N ou $(n_i - r_i)/(N - R)$. Quando documentos não relevantes são recuperados pela consulta inicial, a melhor estimativa para p_i, a probabilidade que um termo ocorra em um documento relevante é simplesmente a probabilidade de sua ocorrência na coleção completa (Salton & Buckley, 1997). Neste caso, $p_i = n_i/N$.

$$p'_i = Pr(x_i = 1|rel)\frac{r_i + n_i/N}{R + 1} \tag{3.16}$$

$$u'_i = Pr(x_i = 1|nonrel)\frac{n_i - r_i + n_i/N}{N - R + 1} \tag{3.17}$$

O fator de ajuste (n_i/N) utilizados nas equações 3.16 e 3.17, substitui o fator 0,5 presentes nas equações 3.14 e 3.15. Quando os documentos relevantes que não foram recuperados for pequeno, podemos utilizar o fator de ajuste alternativo (n_i - r*i*)/(N - R) (Salton & Buckley, 1997).

Salton e Buckley (1997), apontam como vantagem do modelo de realimentação probabilística, a utilização do processo de realimentação ser diretamente relacionado à derivação de um peso para termos da consulta. Ao analisarmos a função de similaridade da equação 3.9, podemos observar que o fator de pesagem de $\log[p_i(1-u_i)/u_i(1-p_i)]$ é aumentada para cada termo

da consulta *i*. Onde é combinado um documento, e o peso do termo ideal sob as condições assumidas de independência do termo e indexação binária do documento (Salton & Buckley, 1997).

3.2.4 Pseudo Realimentação de Relevantes

Pseudo Realimentação de Relevantes (PRR), provem um método para uma análise automática local. Esta técnica automatiza a parte manual da realimentação de relevantes, de modo que o usuário visa melhorar o desempenho da recuperação diminuindo a interação com o sistema (Manning et al., 2008). O método de pseudo realimentação de relevantes consiste em realizar uma recuperação de informação normal para encontrar um conjunto inicial de documentos relevantes, após a recuperação inicial este método assume que os *n* documentos recuperados ordenados no topo da lista de documentos recuperados são relevantes. De posse destes *n* documentos recuperados preliminarmente, o método realimenta a consulta original com as novas informações. Manning et al. (2008) apontam que o método de pseudo realimentação de relevantes tende a funcionar melhor do que o método de análise global, apresentado na Subseção 3.1.2.

Na Figura 8, podemos observar o processo de pseudo realimentação de relevantes. Pseudo realimentação de relevantes é uma variável da realimentação de relevantes que pode lançar mão de todas as técnicas utilizadas na RR.

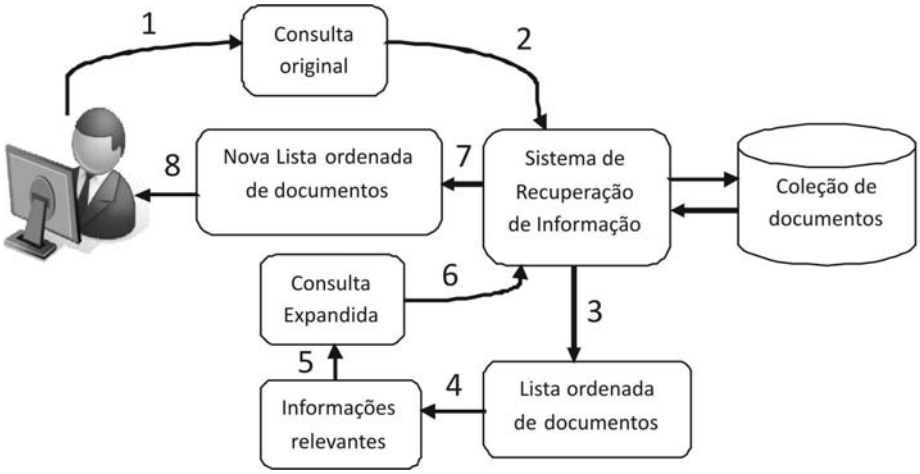

Figura 8 – Processo de Pseudo Realimentação de Relevantes

3.3 Considerações sobre o capítulo

No Capítulo 3 apresentamos uma revisão da literatura sobre as Expansão de Consultas e Realimentação de Relevantes 3.2. Com este estudo apresentamos as duas técnicas que a literatura aponta para a análise de documentos, sendo: (i) análise automática local e (ii) análise automática global. Ao estudarmos EC pudemos ter uma visão mais ampla de suas características, benefícios e limitações, o que foi muito importante para o desenvolvimento da dissertação.

Também neste capítulo, apresentamos um estudo sobre Realimentação de Relevantes (Subseção 3.2), onde abordamos a utilização da RR empregando os modelos clássicos de RI apresentados na Seção 2.1. Além de abordarmos RR, também apresentamos uma variação de RR denominada Pseudo Realimentação de Relevantes na Subseção 3.2.4. O estudo tanto sobre RR como PRR foi fundamental para a definição e aplicação destas técnicas no desenvolvimento desta dissertação.

No próximo Capítulo (Capítulo 4), apresentaremos o modelo de recuperação de informação utilizado nesta dissertação denominado TR+. Apresentaremos detalhadamente as características do Modelo TR+ que foram utilizadas por nós neste trabalho. Na Seção 4.1 apresentamos o processo de Nominalização dos termos, na Seção 4.2 apresetamos a definição das relações lexicais binárias, na Seção 4.3 apresentamos o Conceito de Evidência, na Seção 4.4 apresentamos a formulação da consulta pelo Modelo TR+.

4 O Modelo TR+

Segundo Gonzalez (2005), podemos definir o Modelo TR+ como um modelo de RI, pois além do Modelo TR+ ser capaz de representar textos, também apresenta outras características de um modelo de RI como: indexação, consulta e recuperação de documentos. Uma outra característica que está presente no Modelo TR+ e que o caracteriza como um modelo de RI é a definição do espaço dos descritores. A definição do espaço dos descritores se dá da seguinte forma:

Dados uma coleção $D = \{d\}$, um conjunto de termos $T = \{t\}$ e um conjunto de relacionamentos $R = \{r\}$, um espaço de descritores I_{TR+} é definido como:

$$I_{TR+} = T \cup R$$

onde:

$T = \{t, t = n_1(p)$ ou $t = n_2(p)$ com peso $W_{t,d}$ em $d \in D\}$;

$R = \{r, r = id(t_1 \in T, t_2 \in T)$ com peso $W_{r,d}$ em $d \in D\}$;

p é uma palavra (adjetivo, verbo(incluindo o particípio) ou advérbio) contida em d;

n_1 e n_2 são operações de nominalização;

$id(t_1 \ e \ t_2)$ é uma relação lexical binária (RLB); e

$W_{t,d}$ e $W_{r,d}$ são mensurados utilizando a fórmula baseada no conceito de evidência.

Na Figura 9 proveniente de Gonzalez (2005), podemos observar a fase de indexação de documentos do Modelo TR+. Na Figura 10, também proveniente de Gonzalez (2005), é apresentada a etapa de busca dos documentos no Modelo TR+. Na Figura 9 são apresentadas as etapas necessárias para serem gerados os espaços dos descritores.

O Modelo TR+ utiliza um tratamento equivalente à construção dos descritores de conceitos (termos simples e compostos) tanto para os documentos como para as consultas realizadas em linguagem natural. As consultas são reformuladas com a inclusão de operadores *booleanos* ((i)disjunção lógica entre as RLBs e (ii) conjunção lógica entre os termos e as RLBs) antes do processo de busca e classificação dos documentos. Os relacionamentos entre os termos nominalizados que possuem a capacidade de capturar mecanismos de coesão frásica são chamados de Relações Lexicais Binárias (RLBs) (Gonzalez, 2005).

O primeiro passo proposto no Modelo TR+ é o pré-processamento do texto, utilizando os métodos de tokenização e etiquetagem morfológica. Após o pré-processamento do texto, é realizada a nominalização dos termos. Definindo-se com isso os termos que constituirão os descritores. Descritores são utilizados para descrever conceitos com maior ou menor representatividade (representada por pesos) dentro de um documento textual (Gonzalez, 2005). Finali-

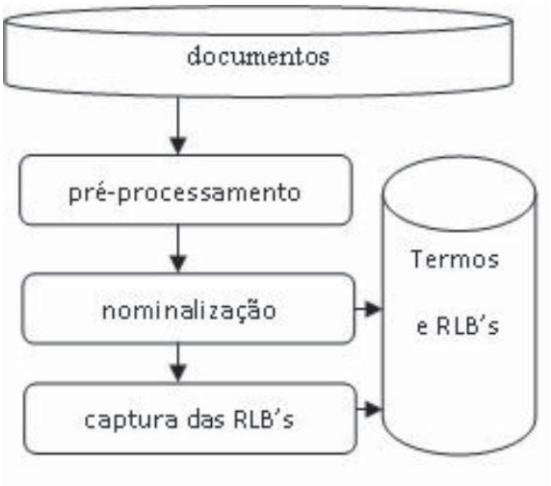

Figura 9 – Fase de Indexação dos documentos no Modelo TR+ (Gonzalez, 2005)

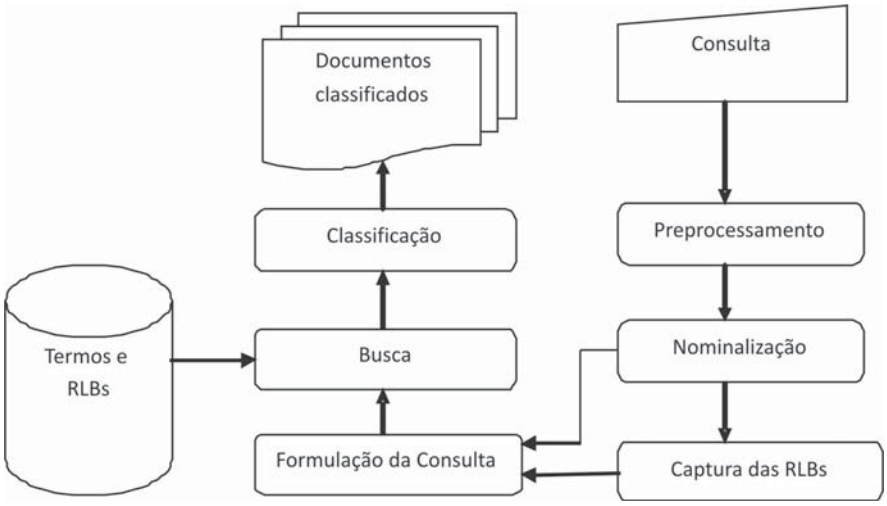

Figura 10 – Fase de Busca dos documentos no Modelo TR+ (Gonzalez, 2005)

zando com a identificação das relações lexicais binárias (RLB), as quais propiciam, ao Modelo TR+, uma forma própria de representação.

4.1 Processo de Nominalização

Gonzalez cita como definição de nominalização Kehdi (2000), que diz que nominalização é um processo de formação de palavras no qual um novo substantivo é derivado de uma palavra existente no léxico, principalmente verbos e adjetivos. Já no Modelo TR+, Gonzalez (2005) define nominalização como sendo, a transformação de uma palavra (adjetivo, verbo (incluindo o particípio) ou advérbio), existente no texto, em um substantivo semanticamente correspondente, formado através de regras válidas de formação de palavras.

No Modelo TR+, substantivos abstratos e concretos são derivados, sendo os substantivos

abstratos representando: eventos, qualidades, estados, ou outras entidades abstratas capazes de serem derivadas de adjetivos, verbos (incluindo particípio) ou advérbios. Já substantivos concretos via de regra representam palavras derivadas de verbos ou de adjetivos. Na Tabela 4, apresentamos exemplos de nominalização, onde substantivos abstratos e concretos são gerados apartir de verbos (incluindo-se particípios), já adjetivos e advérbios quando passados pelo processo de nominalização originam substantivos abstratos.

Tabela 4 – Processo de Nominalização

palavra original	classe	substantivo abstrato	substantivo concreto
correr	verbo	corrida	corredor
soldado	particípio	solda	soldador
sujo	adjetivo	sujeira	-
velho	adjetivo	velhice	velho
oval	adjetivo	-	-
fluvial	adjetivo	-	rio

Utilizando o exemplo apresentado por Gonzalez para exemplificar o processo de nominalização de palavras das quais não são derivadas de substantivos concretos nem de substantivos abstratos, podemos observar o adjetivo ”ovo” (Tabela 4). Comparando-o com o adjetivo ”fluvial”, podemos explicar com mais facilidade a não possibilidade do processo de nominalização. Podemos observar que a equivalência entre ”barco fluvial” e ”barco de rio” não está presente em ”barco oval” e ”barco de ovo”. Isto ocorre porque ”oval” e ”ovo” não são equivalentes, uma vez que ”oval” está relacionato ao seu formato (neste contexto, em forma de ovo). Assim, o adjetivo ”oval” é considerado um descritor deste contexto mais adequado para tal do que o substantivo ”ovo”.

No processo de nominalização, no contexto lexical, existem palavras que se mantêm idênticas após o processo. Na Tabela 4, podemos observar a palavra ”velho”, que pode aparecer tanto na forma de adjetivo quanto como substantivo (neste caso, substantivo concreto).

No Modelo TR+, os termos adicionados no espaço de descritores, são substantivos originários de um dado texto ou ainda palavras de outras classes, quando não há a possibilidade de se nominalizá-las. Os substantivos originários de um dado texto são normalizados utilizando o processo de lematização. São adicionados no espaço de descritores substantivos oriundos da nominalização de adjetivos, advérbios e verbos (incluindo-se particípios). A concepção dos termos nominalizados desconsidera o tratamento de acentuação ou a forma maiúscula ou minúscula dos termos, sendo esses definidos sem acento e na forma minúscula. Esses termos denonimados de ”termos nominalizados” constituem os argumentos das RLBs. Apresentaremos em mais detalhes as RLBs na Seção 4.2.

4.2 Definições das Relações Lexicais Binárias

Relações lexicais binárias (RLBs) são relacionamentos entre termos nominalizados, que capturam mecanismos de coesão frásica (Gonzalez, 2005). O processo de identificação dos pesos dos descritores pode utilizar-se das relações lexicais binárias para tal.

Segundo Gonzalez, podemos categorizar as RLBs em três tipos (classificação, restrição e associação) para representar o texto. Cada tipo de RLB possui um formato próprio, apresentando argumentos específicos que desempenham papéis próprios. A forma de apresentação de uma RLB é dada da seguinte maneira:

$$id(t_1, t_2) \tag{4.1}$$

Onde *id* é um identificador de relação; t_1 e t_2 são argumentos (constituidos de termos nominalizados). Podemos representar os tipos de RLBs e suas características da seguinte maneira:

- **Classificação**: o identificador da relação é representado pelo sinal de igualdade *(=)*; t_1 representa uma subclasse ou uma instância de t_2; e t_2 representa uma classe.

  Exemplo:

  =(gato,animal)
  =(miau,animal)

- **Restrição**: o identificador da RLB é representado por uma preposição; t_1 representa um elemento modificado; t_2 representa um elemento modificador.

  Exemplo:

  de(equipe,futebol)
  com(analista,experiencia)

- **Associação**: o identificador da RLB é representado por um evento; t_1 é um sujeito; t_2 é um objeto (direto ou indireto) ou adjunto.

  Exemplo:

  superação(trabalhador,dificuldade)
  moradia.em(rainha,londres)

Nos exemplos apresentados para as RLBs do tipo associação, podemos observar que seu *id*, pode se apresentar na forma de preposição, desta forma fica garantido que a relação apareça com dois argumentos. Segundo Gonzalez, as RLBs dos tipos restrição e associação podem (mas não somente) apresentar em seu *id* o seguinte formato:

evento.preposicao

No exemplo exposto à RLB do tipo restrição "*de(equipe,manutencao)*", podemos admitir a utilização para o *id* do formato *evento.preposicao* teríamos então:

é.de(equipe,manutencao)

No Modelo TR+, o cálculo do peso dos descritores é realizado utilizando-se as RLBs, não havendo distinção quanto ao seu tipo. Para o Modelo TR+ a identificação do tipo das RLBs é importante para a organização destas em arquivos de índice, visando com isso o aumento do desempenho da pesquisa das RLBs durante o processo de busca na recuperação dos documentos.

Gonzalez define as relações das RLBs como sendo assimétricas, pelo fato dos seus argumentos possuírem papéis específicos, constituindo assim uma estrutura de relacionamentos capacitado para a representação dos textos contidos nos documentos (ver exemplo transcrito de Gonzalez (2005) no Anexo B).

A classificação das RLBs pode ser realizada tanto pelo seu tipo (classificação, restrição e associação) quanto pela sua nominalização (Original ou Derivada). Podemos exemplificar a classificação das RLBs quanto a sua nominalização da seguinte forma:

- **Original**:quando nenhum componente sofreu nominalização
  Exemplos:

movimento da rua - de(movimento,rua)
rapidez da jogada - de(rapidez,jogada)
atuação do jogador - de(atuacao,jogador)
corredor João - =(corredor,joao)

- **Derivada**: quando ao menos um compontente sofreu nominalização
  Exemplos:

rua calma - de(calma,rua)
jogou rapidamente - de(rapidez,jogada)
jogador atuou - de(atuacao,jogador)
João correu - =(joao,corredor)

O objetivo apontado por Gonzalez na inclusão das RLBs no espaço dos descritores é de aumentar a amplitude da descrição dos textos. Gonzalez ao descrever as RLBs afirma que estas apresentam relações semânticas equivalentes as apresentadas na estrutura *Qualia* na teoria do Léxico Gerativo, justificando com assim a utilização das RLBs na recuperação de informação.

A estrutura *Qualia*, na teoria do Léxico Gerativo, descreve um item lexical 'α' através de quatro papéis: formal, constitutivo, agentivo e télico.

Podemos exemplificar tais papéis da seguinte maneira:

- **Formal**: diferencia 'α' em um amplo domínio;
  Exemplo:

  =(carro,maquina)

  O carro seria distinguido como uma máquina.

- **Constitutivo**: indica o que faz parte de 'α';
  Exemplo:

  de(carro,aluminio)

  Neste exemplo o carro seria fabricado de alumínio.

- **Agentivo**: especifica qual a razão de 'α' passar a existir;
  Exemplo:

  por(composição,autor)

  Neste exemplo é especificado que a composição é referente a um dado autor.

- **Télico**: explica qual a função ou propósito de α;
  Exemplo:

  conserto(mecanico,vazamento)

  Neste exemplo a função do mecânico seria consertar o vazamento.

Apesar destas características, Gonzalez afirma que não pretende que as RLBs "interpretem" com distinções, indicações, especificações ou explicações dos tipos apresentados acima. O objetivo é caracterizar as RLBs como descritores de tais fatos sem a necessidade da utilização de rótulos, excetuando-se a RLB do tipo classificação que utiliza inevitavelmente o rótulo "=".

Sem a presença de rótulos pré-estabelecidos, Gonzalez afirma que não tem como pretenção que a preposição ”para” indique um propósito, ou ainda que a preposição ”por” represente uma ação realizada pro algum agente. A definição das preposições nas RLBs é dada através da utilização de regras de identificação (ver Anexo A transcrito de Gonzalez (2005)), e tem como objetivo, descrever algo que o sistema não consegue interpretar, mas entretanto, mesmo que seja representado em ocorrência sintática distintas, deva ser descrito da mesma forma. Com tudo isso, Gonzalez pretende com a utilização das RLBs, que estas sejam capazes de descrever conceitos independetemente da forma como estes aparecem no texto. Assim pode-se dizer que as RLBs conceitualizam a ”evidência” aplicadas aos descritores. O conceito de evidência apresentaremos a seguir na Seção 4.3.

4.3 Conceito de Evidência

Segundo Ferreira (1999) e Houaiss (2002), evidência é a condição do que se destaca, é a qualidade do que é evidente e, por sua vez, evidente é aquilo que não oferece ou não dá margem à dúvida. No Modelo TR+ o cálculo dos pesos dos descritores é realizado utilizando o conceito de evidência. Para o cálculo do peso dos descritores além da freqüência de ocorrência do descritor no texto, o Modelo TR+ também utiliza a ocorrência das RLBs.

Gonzalez (2005) esclarece que o resultado do cálculo do peso de um descritor no Modelo TR+ leva em consideração (i) o processo de nominalização, (ii) a capacidade das regras para identificação de RLBs de deduzir estruturas de dependência evidentes e (iii) a formulação do cálculo do peso dos descritores.

A representatividade dos descritores é impactada pela nominalização dos mesmos, já que este processo de normalização lexical coloca em um único descritor diferentes palavras. Descritores que passam pelo processo de normalização lexical tendem a possuir um peso maior quando comparados com aqueles que não sofreram tal processo, pois os descritores normalizados acumulam a freqüência de ocorrência de outros descritores, uma vez que este representa um conjunto de palavras. Gonzalez destaca que este processo de normalização pode ser incluído no conceito de transformação de termos ”ruins” em termos ”bons” apresentado por Salton e MacGill (1983).

A definição da representatividade dos descritores também é impactada pelas regras utilizadas para a identificação das RLBs, pois tais regras são capazes de reconhecer somente estruturas de dependência evidentes. As dependências com preposições a direita após a segunda preposição, são tratadas como sendo ”não evidentes”. Gonzalez (2005) apresenta o seguinte exemplo:

(a) ”arrombamento do cofre com explosivos”
(b) ”arrombamento do cofre com jóias”

Analizando estes exemplos, somente a RLB ”*de(arrombamento,cofre)*” seria identificada. Já

as RLBs, "*com(arrombamento,explosivo)*", "*com(cofre,explosivo)*", "*com(arrombamento,joia)*" e "*com(cofre,joia)*" não são reconhecidas. Com isso alguns descritores perdem representatividade (RLBs e os termos nelas presentes como argumento) sendo penalizados pois não atende ao conceito de evidência (onde deve haver destaque e não pode haver dúvidas). Gonzalez destaca que essa abordagem possui duas vantagens, (i) menor esforço computacional (pelo não tratamento de ambiguidades), (ii) a utilização do conceito de evidência na influência do cálculo do peso dos descritores faz com que quanto mais evidente mais representativo o descritor será.

Considerando o exemplo apresentado por Gonzalez, *(a) "arrombamento do cofre com explosivos"*, podemos apresentar o cálculo do grau de representatividade através das seguintes descrições: (i) dos conceitos relacionados a cada um de seus argumentos e (ii) a descrição de seus relacionamentos. Seguindo o mesmos exemplo, *"arrombamento do cofre com explosivos"*, a RLB deve receber 3 unidades de evidência (1 unidade pois há um *"arrombamento"*; 1 unidade pois há um *"cofre"*; e 1 unidade por haver um *"arrombamento do cofre"*). Cada ocorrência dos descritores "*arrombamento*" e "*cofre*" receberia $1\frac{1}{2}$ unidade, que é metade do valor atribuído à RLB. Por ultimo, o descritor "explosivo", por ser o menos evidente, receberá $\frac{1}{2}$ unidade de evidência, diminuida de 1 unidade pela falta de coesão evidente. O mesmo ocorre "joia" no exemplo (b). O descritor envolvido recebe uma unidade de evidência a cada nova coesão.

Gonzalez apresenta a seguinte explicação para o cálculo do grau de representatividade dos descritores: "os termos t_1 e t_2 e a RLB *r*, encontrados em uma consulta *q*, têm dupla contribuição no cálculo do valor de relevância de um documento *d*, caso t_1 e t_2 estejam relacionados através de *r* em *d*. Do contrário, se t_1 e t_2 ocorrem em *d* mas não estão relacionados através de *r*, a contribuição é simples e, assim, *d* tende a perder posições na classificação por relevância a *q*".

4.3.1 Cálculo do peso dos Descritores e do valor de Relevância

Para o cálculo do peso dos descritores o Modelo TR+ utiliza a abordagem probabilística (Baeza-Yates & Ribeiro-Netto, 1999), pois tal abordagem mostrou-se mais eficiente para a recuperação de informação, segundo Gonzalez (2005). Entretanto, Gonzalez deixa claro que o Modelo TR+ pode se utilizar da abordagem vetorial (Baeza-Yates & Ribeiro-Netto, 1999) para o cálculo dos pesos dos descritores.

A Equação 4.2, uma adaptação da fórmula OKAPI BM25 apresentada na Equação 4.3 sem o fator *IDF* (Gonzalez atesta que a utilização do *IDF* não apresentou melhoria nos resultados dos experimentos), é adotada pelo Modelo TR+. O peso $W_{i,d}$ do descritor *i* no documento *d* é dado por:

$$W_{i,d} = \frac{w_{i,d}(K_1 + 1)}{K_1((1 - b) + b\frac{DL_d}{AVDL}) + w_{i,d}} \tag{4.2}$$

onde:

- $w_{i,d}$ é a freqüência do descritor *i* no documento *d*;
- K_i, *b*, DL_d e *ADVL* são os mesmos componentes utilizados na fórmula Okapi BM25.

$$W_{i,d} = \frac{w_{i,d}(k_1 + 1)}{k_1((1-b) + b\frac{DL_b}{AVDL} + w_{i,d})} IDF_i \tag{4.3}$$

onde:

- $w_{i,d} = f_{i,d}$ é a freqüência de ocorrência de *i* em *d*;
- K_i e *b* são parâmetros;
- DL_d é o comprimento (a quantidade de palavras) do documento *d*;
- *AVDL* é o comprimento médio dos documentos da coleção;
- IDF_i = Log $\frac{N}{df_i}$;
- *N* é o número de documentos na coleção;
- df_io número de documentos onde *i* ocorre.

A evidência $w_{i,d}$, representada através de $w_{t,d}$ para um termo *t* em um documento *d*, é calculada da seguinte forma no Modelo TR+:

$$w_{t,d} = \frac{f_{t,d}}{2} + \sum_r f_{r,t,d} \tag{4.4}$$

onde:

- $f_{t,d}$ é a freqüência de ocorrência de *t* em *d* e
- $f_{r,t,d}$ é a quantidade de RLBs onde *t* é argumento em *d*,

e para uma RLB *r*, a evidência $w_{i,d}$ em um documento *d*, representada por $w_{r,d}$ é:

$$w_{r,d} = f_{r,d}(w_{t1,d} + w_{t2,d}) \tag{4.5}$$

onde:

- $f_{r,d}$ é a freqüência de ocorrência de *r* em *d* e
- $w_{t,d}$ é a evidência do argumento *d* de *r* em *d*;

O Anexo B, transcrito de Gonzalez (2005), apresenta alguns exemplos onde é apontado resultados do cálculo baseado em evidência, em comparação à formulação baseada apenas em freqüência de ocorrência. Os termos e a RLBs são obtidos e têm seus pesos calculados utilizando a mesma abordagem tanto para uma consulta *q*, quanto para os documentos. Entretanto, para cada RLB *r* presente na consulta *q*, sendo, *r = id(*t_1*,*t_2*)*, uma RLB *r'*é incluida na consulta Booleana *qb*, sendo, *r' = id'(*t_1*,*t_2*)*, onde *id'* é qualquer identificador diferente de *id* (conforme é exemplificado na Seção 4.4). O peso $W_{r',q}$ de *r'* depende do peso $W_{r,q}$ de *r*, sendo penalizado por possuir identificador diferente, mesmo que *r* e *r'* possuam os mesmos argumentos. $W_{r'}$, *q* é dado por:

$$W_{r',q} = \frac{W_{r,q}}{2} \tag{4.6}$$

Para se obter o valor de relevância $VR_{d,q}$ tanto para um documento *d* como para uma consulta *q*é utilizada a seguinte equação:

$$VR_{d,q} = \sum_{i} (W_{i,d}, W_{i,q}) \tag{4.7}$$

onde:

- $W_{i,d}$ é o peso de termos e/ou RLBs do documento *d* e
- $W_{i,q}$ é o peso de termos e/ou RLBs da consulta *q*.

Após a definição dos termos e RLBs, assim como seus respectivos pesos, os documentos tem sua classificação dependente do valor da relevância dos mesmos e da formulação da consulta Booleana (Gonzalez, 2005).

4.4 Consulta Booleana

A formulação da consulta Booleana *qb*, no contexto do Modelo TR+, é realizada de acordo com a gramática apresentada a seguir, com formalismo BNF (Gonzalez, 2005):

qb→ *disjunção*RLBs *OU* (*conjunção dos Termos*)
*disjunção*RLBs → *r OU disjunção*RLBs | ϵ
conjunçãoTermos → (*disjunçãoTermos*) *E conjunçãoTermos* | ϵ
disjunçãoTermos → n_1*(p) OU* n_2*(p)*
r → RLB
p → adjetivo | advérbio | particípio | substantivo | verbo
OU → operador Booleano de disjunção
E → operador Booleano de conjunção
ϵ → elemento vazio

Neste esquema, a relação entre um descritor X e ϵ, possui o mesmo valor de X. A consulta "*pintura restaurada*", então será formulada no Modelo TR+, da seguinte maneira (Gonzalez, 2005):

$$r1\ OU\ r2\ OU\ ((n_1(\text{p1})\ E\ n_2(\text{p2})\ OU\ n_2(\text{p2})))$$

onde:

- *r1* = de(restauracao,pintura),
- *r2* = *r1'* = $\neq$de(restauracao,pintura)),
- *n1(p1)* = ϵ,
- *n2(p1)* = pintura,
- *n1(p2)* = restauracao,
- *n2(p2)* = restaurador,
- *p1* = pintura e
- *p2* = restaurada.

A notação "$\neq$de" significa qualquer identificador diferente de "de".
Os documentos recuperados são, então, classificados em dois grupos:

- (i)grupo superior, documentos de maior relevância: estão presentes neste grupo, documentos que possuem pelo menos uma das RLBs da consulta ou, possuem todos os termos da consulta;
- (ii)grupo inferior, documentos de menor relevância: estão presentes neste grupo, documentos que possuem pelo menos um dos termos da consulta.

4.5 Considerações sobre o capítulo

No Capítulo 4 apresentamos as características que definem o Modelo TR+ como um modelo de RI capaz de indexar e recuperar documentos. Na Seção 4.1 apresentamos o processo de transformação de uma palavra do texto em um substantivo semanticamente correspondente, chama de Nominalização (Gonzalez, 2005). Na Seção 4.2 apresentamos o processo de definição das RLBs, que são os relacionamentos entre os termos nominalizados. Na Seção 4.3, apresentamos o conceito de Evidência que é utilizado pelo Modelo TR+ para o cálculo do peso dos descritores. Na Seção 4.4, apresentamos a formulação da consulta no Modelo TR+. A consulta

no Modelo TR+ utiliza operadores booleano para a sua construção. O estudo do Modelo TR+ foi parte fundamental, uma vez que o modelo é parte central dessa dissertação. O estudo do Modelo TR+ nos possibilitou colocar em prática os experimentos de aplicação da RR e PRR, uma vez que nos deu embasamento para o entendimento das características do modelo.

No próximo capítulo, Capítulo 5, apresentaremos alguns trabalhos selecionados por nós entre os estudados durante o desenvolvimento da dissertação. Os trabalhos foram escolhidos por colaborarem para a conclusão da dissertação, Na Seção 5.1 apresentamos a proposta desenvolvida por (Custis & Al-Kofahi, 2007) de avaliação entre os termos da consulta e dos documentos para expansão de consultas. Na Seção 5.2, apresentamos o trabalho desenvolvido por Vechtomova e Karamuftuoglu (2007) que utilizam a análise lexical para a seleção dos termos que farão parte do processo de EC. Na Seção 5.3, apresentamos o trabalho desenvolvido por Chirita e Nejdl (2007), que busca a personalização da EC para a recuperação de documentos *Web* utilizando informações presentes nos computadores dos usuários. Na Seção 5.4, apresentamos o trabalho desenvolvido por Orengo e Huyck (2006), onde através da RR os autores buscam recuperar documentos utilizando informações multilíngües. Na Seção 5.5, apresentamos o trabalho desenvolvido por Lee et al. (2008), e que visa a extração de novas amostragens de termos para EC baseados em agrupamentos utilizando PRR.

5 Trabalhos Correlatos

Nesta seção apresentaremos alguns trabalhos que se utilizam de expansão de consulta e que nortearam o desenvolvimento do trabalho até o presente momento.

5.1 Uma Nova Proposta para Avaliação de Expansão de Consulta: Má Combinação entre termos da consulta e dos documentos (Custis & Al-Kofahi, 2007)

Custis e Al-Kofahi (2007) não apresentam exatamente um método de EC, mas sim uma proposta para a avaliação da expansão de consultas utilizando a combinação dos termos da consulta efetuada pelos usuários e os termos presentes nos documentos em uma coleção de um domínio específico. Os termos utilizados na consulta são retirados dos documentos julgados relevantes um a um, possibilitando com isso determinar a eficiência de diferentes sistemas de recuperação de informação no que diz respeito à perda desses termos.

Para a validação da proposta apresentada pelos autores, foram realizados quatro experimentos: (i) dois experimentos com a utilização da fórmula OKAPI para o cálculo dos pesos dos termos (Huang et al., 2006) (com e sem o uso de pseudo realimentação de relevantes para a expansão de consulta); (ii) um experimento fazendo uso do mecanismo de busca proprietário TCS, Thomson Concept Search; e (iii) um experimento utilizando o modelo de linguagem de consulta probabilística (Query Likelihood)(Zhou & Croft, 2005). O TCS utiliza um corpus externo como fonte de conhecimento tematicamente relacionado à coleção de documentos que será pesquisada.

Para a validação dos experimentos foram utilizadas duas coleções de teste para os quatro sistemas de recuperação de informação já mencionados. As duas coleções de teste utilizadas são: o TREC AP89 (TIPSTER disco 1), que é uma coleção de textos da *Text Retrieval Conference*, e a coleção proprietária de documentos de casos legais chamada *FSupp*.

Nos experimentos realizados, a estratégia escolhida de remoção dos termos da consulta para toda a coleção de documentos fez uso do *Inverse Document Frequence* (IDF) (Salton & MacGill, 1983). Termos com alto valor para o IDF influenciam a classificação dos documentos. Termos com alto valor para IDF geralmente são termos do domínio específico, que são menos comuns, sendo difícil para uma pessoa não especialista reconhecê-los. Por esse motivo a remoção desses termos com alto valor para o IDF são removidos em primeiro lugar. Para comparar

a eficiência de cada sistema de recuperação de informação foram utilizadas MAP com precisão para dez documentos (P@10) (Turpin & Scholer, 2006), e abrangência para mil documentos.

Os autores concluem que a sua proposta de avaliação de sistemas de recuperação de informação permite medir o grau de melhoria (ou não) da combinação de termos entre a consulta e documentos considerados relevantes. A avaliação dos sistemas de recuperação de informação é realizada utilizando somente coleções inteiras de documentos evitando, com isso o uso na expansão de consulta de uma combinação de termos que não resulte em uma recuperação de documentos eficiente para as necessidades dos usuários. Outra contribuição importante é que os resultados podem ser avaliados independentemente das métricas escolhida para tal. Também, os autores mostram que é possível modelar o comportamento de usuários analisando a combinação de termos que estes utilizam na consulta em dois grupos: usuários especialistas e usuários iniciantes.

Com o estudo do trabalho realizado por Custis e Al-Kofahi (2007), identificamos a importância e viabilidade de uso da técnica de EC Pseudo Feedback para a aplicação em conjunto com o Modelo TR+ na recuperação de informação. Unido a isso,outra importante contribuição do trabalho apresentado por Custis e Al-Kofahi (2007) foi trazer a oportunidade de um melhor conhecimento de uma situação de uso da fórmula OKAPI.

5.2 Expansão de Consulta com termos selecionados usando análise da coesão lexical dos documentos (Vechtomova & Karamuftuoglu, 2007)

Vechtomova e Karamuftuoglu (2007) apresentam uma proposta para expansão de consultas utilizando ligações lexicais coesivas entre os termos da consulta e os termos dos documentos vizinhos aos termos da consulta no documento. Partes do texto (*Snippets*) vizinhas ao termo da consulta dentro do documento são avaliadas para expansão de consultas de forma automática.

No trabalho apresentado é explorada a eficácia da utilização de *snippets* para se expandir consultas de forma interativa com o usuário. Os autores comparam expansão de consulta utilizando *snippets* do texto e expansão de consulta com o uso de documentos inteiros. Também é mencionada no trabalho uma comparação de expansão de consultas utilizando partes do texto selecionado pelo usuário versus a expansão de consulta com utilização de documentos inteiros julgados relevantes pelo usuário.

A avaliação foi conduzida no TREC 2005 (*Text Retrieval Conference*), considerando o uso de termos de ligação e termos vizinhos de partes do texto em comparação com termos selecionados de textos inteiros. A proposta apresentada pelos autores foi comparada com a expansão de consulta utilizando a freqüência dos termos no documento como peso, onde todos os termos são extraídos de um texto completo de um documento reconhecidamente relevante, e ordenado.

Os autores apresentaram experimentos com expansão de consultas utilizando pseudo reali-

mentação de relevantes para avaliar a proposta de expansão de consulta sem retorno de relevância. Ao término dos estudos concluem que ao apresentarem aos usuários os termos de partes do texto dentro do contexto como auxílio para a expansão da consulta, os usuários selecionam termos mais eficientes, em contrapartida ao que ocorre quando expõem aos usuários termos fora desse contexto.

Os autores ainda finalizam constatando que existe uma significativa diferença no número de ligações lexicais entre termos de consultas distintas em conjuntos de documentos relevantes quando compardados a um conjunto de documentos não relevantes.

O trabalho apresentado por Vechtomova e Karamuftuoglu (2007) foi de grande valor para a formulação de nossa proposta, pois ofereceu uma visão prática da utilização da técnica de EC pseudo realimentação de relevantes, chamada neste trabalho por Vechtomova e Karamuftuoglu (2007) de *Blind Feedback*. O trabalho também nos apresentou a utilização de *snippets* dos documentos para a EC, a utilização de *snippets* associada ao Modelo TR+, é uma alternativa à proposta apresentada nesta dissertação. Entretanto devido a necessidade de modificações no Modelo TR+, não lançaremos mão de tal abordagem. O trabalho apresentado por Vechtomova e Karamuftuoglu (2007), fortaleceu a utilização da PRR como uma técnica de EC a ser aplicada junto ao Modelo TR+.

5.3 Expansão de Consulta Personalizada para a Web (Chirita & Nejdl, 2007)

Chirita e Nejdl (2007) propõem melhorar o resultado da recuperação produzido por consultas em um ambiente *Web* expandindo-as com termos extraídos de dados obtidos a partir do computador do usuário, em um assim denominado Repositório de Informações Pessoais, dessa forma personalizando o resultado da busca.

Os autores introduziram cinco técnicas para geração de palavras-chave em uma consulta, analisando dados do usuário. Para o ajuste dos termos e análise do nível da co-ocorrência dos componentes, são utilizados *thesauri* externos. A expansão da consulta do usuário é feita com novos termos usualmente extraídos de um grande *thesaurus*, como a *WordNet* (Wordnet, 2008), de onde se extrai os sinônimos dos termos da consulta original.

Ainda, os autores apresentam análises sob quatro cenários diferentes, mostrando que algumas destas propostas aumentam o desempenho especialmente em consultas ambíguas, melhorando a qualidade da classificação da resposta.

Chirita e Nejdl (2007), utilizam expansão de termos baseados na *WordNet*, entretanto os autores basearam seus estudos na análise do relacionamento entre os dados do repositório pessoal e os termos da consulta original para definir novas palavras que formarão a consulta expandida.

A extração se dá através do cálculo da freqüência do termo nos documentos do repositório

pessoal. A análise local do repositório pessoal está diretamente relacionada com a utilização de Pseudo Realimentação de Relevantes, buscando uma melhora na geração de palavras-chave utilizadas na expansão de consultas que buscam informações na *Web*.

De um conjunto de documentos classificados como os mais relevantes são extraídos partes do texto que melhor os representam. Seguindo a etapa de análise, os autores apresentam um estudo sobre a composição lexical para a identificação automática dos conceitos mais importantes na coleção de documentos. Os autores utilizaram para determinar a composição lexical o processo de análise de substantivos, verificando a combinação dos documentos do repositório pessoal do usuário para todas as composições.

Após a análise local, os autores dirigiram seus estudos para a análise global, para extrair informações inferindo novos termos para a expansão da consulta. Para a análise global foram estudadas duas técnicas, a primeira baseada na estatística da co-ocorrência de termos e a segunda na expansão baseada em *thesaurus*.

Somente substantivos são utilizados para esta proposta, os autores justificam essa escolha pela grande capacidade que os substantivos apresentam de conter informações conceituais. Os termos dos resumos com maior grau de relação para cada termo da consulta são identificados e finalmente a correlação dos termos resultante é calculada para a consulta inteira. Os autores ainda consideram um passo importante o cálculo do coeficiente de similaridade. A proposta utiliza a freqüência do documento para um dado termo x, e o número de documentos que contenham tal termo x. O número de termos utilizados na expansão de consulta é limitado, para melhorar a pontuação alcançada.

Para a validação da proposta foram realizados alguns experimentos. O primeiro passo dos experimentos foi a instalação do mecanismo de busca proposto nesse trabalho, baseado no Lucene (Broadbent et al., 2006), indexando todos os conteúdos armazenados localmente. A máquina de um único usuário foi alvo dos experimentos, onde foram realizadas diariamente quatro consultas.

Para cada consulta, foram selecionados os cincos endereços *Web* melhor classificados gerados para cada versão do algoritmo. Cada resultado foi ”misturado” em um conjunto de noventa endereços *Web*. Com isso, cada assunto teve que ser acessado entre o conjunto de documentos (325 documentos) para todas as quatro consultas. Foram realizadas 72 consultas e mais de 6.000 endereços *Web* foram avaliados durante o experimento. Para cada endereço *Web*, o testador teve que entregar uma avaliação com os valores sendo: 0 (não relevante), 1 (relevante) e 2(altamente relevantes). Finalmente, a qualidade de cada classificação é estimada usando a versão normalizada do Ganho Cumulativo Descontado (DCG) (Järvelin & Kekäläinen, 2000). DCG atribui mais peso aos documentos que foram melhores classificados no processo de recuperação. Analisando os resultados dos quatro cenários, o melhor cenário obteve um aumento de desempenho no que tange a qualidade dos documentos recuperados de 51,21%.

Ao término do estudo do trabalho proposto por Chirita e Nejdl (2007), ficou clara a dificuldade de se aplicar a EC utilizando informações contidas na máquina do usuário ao Modelo

TR+, uma vez que, para o uso dessa referência seria fundamental dispor de um *thesaurus* externo como a *WordNet*. Com isso a aplicação dessa proposta ao Modelo TR+ foi descartada pelo tempo exigido para a construção de um *thesaurus* à língua portuguesa. O trabalho proposto por Chirita e Nejdl (2007) fortaleceu a nossa decisão de utilizar a técnica de EC pseudo realimentação de relevantes, eliminando a participação do usuário para melhorar a qualidade das informações recuperadas junto ao Modelo TR+.

5.4 Realimentação de Relevantes e Recuperação de Informações Multilíngüe (Orengo & Huyck, 2006)

Orengo e Huyck (2006) apresentam um estudo de realimentação de relevantes em um ambiente de recuperação de informação multilíngüe. Os autores apresentam um experimento no qual, falantes nativos em português julgam a relevância de documentos escritos em inglês; documentos traduzidos manualmente para o português e documentos traduzidos de forma automática também para o português. Orengo e Huyck (2006) buscaram como principal objetivo em seus experimentos responder duas questões: (i) quão bom pesquisadores nativos da língua portuguesa são para reconhecer documentos relevantes escritos em língua inglesa, comparando documentos traduzidos de forma manual e automaticamente para o português; e (ii) qual é o impacto do mal julgamento de documentos na realimentação de relevantes. Orengo e Huyck (2006) mostram como resultado para seu experimento que a tradução realizada de forma automática apresentou-se tão eficiente quanto a tradução realizada manualmente. Além disso, o impacto do mau julgamento dos documentos no desempenho da realimentação de relevantes é moderado e varia muito para diferentes tópicos.

Orengo e Huyck (2006) utilizaram para este trabalho o sistema CLIR (Orengo & Huyuck, 2002). O sistema CLIR foi implementado utilizando indexação semântica latente[1] (ISL)(do inglês *Latent Semantic Indexing* - LSI). Segundo Orengo e Huyck (2006), o principal objetivo da utilização da LSI para a implementação do CLIR, foi de prover o sistema de recursos capazes de compara segmentos de textos escritos em uma língua com segmentos de textos escritos em outra língua sem a necessidade da tradução destes textos. Os autores utilizaram o *software* para tradução chamado SYSTRAN 3.0 *Professional*, para traduzir cerca de 20% dos documentos da coleção. Orengo e Huyck (2006), escolheram utilizar o SYSTRAN 3.0 *Professional* pelo fato de ser muito utilizado em trabalhos que se utilizam do sistema CLIR descritos na literatura, e também pelo fato do SYSTRAN 3.0 Professional ser adotado nos experimentos realizados durante o CLEF (*Cross-Language Evaluation Forum*). O corpus utilizado no experimento consiste de mais de 113.000 artigos de notícias do Jornal americano *Los Angeles Times*. Esta coleção foi utilizada por fazer parte dos experimentos realizados no CLEF.

[1] Utilizaremos a sigla originária do inglês LSI ao nos referirmos sobre indexação semântica latente.

Os experimentos foram realizados a responder duas questões principais:

- (i) Quão bom pesquisadores nativos em português são em reconhecer a relevância de documentos escritos em inglês, comparando documentos traduzidos manualmente com documentos traduzidos automaticamente para o português?

- Qual é o impacto do mau julgamento dos documentos na performance que pode ser alcançado com a RR?

Para responder estas duas questões Orengo e Huyck (2006) recrutaram 27 usuários[2] com idade média de 29 anos, falantes nativos da língua portuguesa com conhecimento básico ou sem conhecimento da língua inglesa que não conseguiriam expressar as consultas em língua inglesa mas com bom conhecimento de mecanismos de busca.

Os autores extraíram seis tópicos de consulta de um total de cinqüenta consultas do CLEF do ano de 2002, sendo utilizada uma versão em português dos tópicos. O critério de seleção dos tópicos com base no resultado inicial foi:

- (i) Selecione tópicos com mais de 10 documentos relevantes. Este critério visa prevenir no quais todos os documentos relevantes são apresentados para o usuário por realimentação.

- (ii) Selecione tópicos que tem documentos relevantes entre os 10 primeiros recuperados. Assim o método RR utiliza somente realimentação positiva, este critério foi utilizado para prevenir a situação na qual o usuário não julga qualquer documento como sendo relevante.

Dezessete dos cinqüenta tópicos atenderam as condições impostas pelos autores. Os experimentos se deram com apresentação aos pesquisadores dos tópicos de consulta escritos em português e a uma lista de documentos ordenados retornados em resposta a consulta original. A lista de documentos ordenados foi produzida pela presença de todos os termos do título e descrição utilizando como consulta no sistema CLIR - LSI. Os participantes foram convidados a classificar cada documento em relação ao tópico em uma das três categorias: relevante; não relevante; não sei. Cada participante leu 6 consultas e 10 documentos para cada consulta, atingindo 60 julgamento de relevância por usuário e 1620 no total. Os documentos foram apresentados na forma completa e foram expostos em uma das seguintes formas:

- (i) o texto original em inglês, retornado do sistema CLIR-LSI (sistema 1),

- (ii) uma tradução automática produzida utilizando o software SYSTRAN 3.0 *Professional* (sistema 2),

- (iii) uma tradução humana.

[2] Alunos da Universidade Católica de Pelotas, htpp://www.ucpel.tche.br

Após recolhido o julgamento de todos os 27 pesquisadores, as consultas foram realizadas, submetidas e avaliadas novamente para a precisão e abrangência. RR foi aplicada substituindo a consulta original pela média vetorial dos documentos selecionados pelos usuários como sendo relevantes.

Depois de realizada a avaliação dos resultados, Orengo e Huyck (2006) concluíram que:

- (i) 44% dos participantes mostraram-se hábeis para avaliar documentos em língua inglesa,
- (ii) a tradução automática pode realmente ajudar pesquisadores na avaliação da relevância, apesar de produzir documentos deselegantes para a leitura. Os participantes julgaram com a mesma eficiência documentos traduzidos de forma automática e documentos traduzidos manualmente,
- (iii) existe uma moderada correlação negativa entre documentos julgados erroneamente e a melhoria que pode prover o método de RR,
- (iv) o fator que impacta a mudança no desempenho varia muito de um tópico para outro. Cada tópico responde de forma própria aos julgamentos errados. Porém, as características de cada tópico que determina o relacionamento entre a mudança no desempenho e erros no julgamento permanecem obscuros,
- (v) não foram encontrados relacionamentos entre a mudança no desempenho e a dificuldade dos tópicos ou crença no julgamento ou o conhecimento do assunto,
- (vi) muitos usuários consideraram o sistema CLIR útil e gostariam que os resultados fossem traduzidos em outras línguas.

Com o estudo do trabalho apresentado por (Orengo & Huyck, 2006) pudemos nos familiarizar com o método de realimentação de relevantes, assim como, com a forma de avaliar a relevância dos documentos recuperados pela consulta original. Assim sendo, pudemos desenvolver os experimentos com RR em conjunto com o Modelo TR+ para a RI aplicados nesta dissertação.

5.5 Um método de extração de nova amostragem baseado em grupos para Pseudo Realimentação de Relevantes (Lee et al., 2008)

Neste trabalho Lee et al. (2008) apresentam um método para definir novas amostragens baseada em agrupamentos de documentos, para assim selecionar melhores documentos que serão utilizados para a EC utilizando PRR. A idéia principal desse trabalho é encontrar em uma recuperação inicial, um conjunto de documentos ”dominantes” que serão utilizados para a EC e com isso enfatizarem o tópico central de uma consulta.

Lee et al. (2008) assumem que documentos dominantes para uma consulta, são aqueles que possuem uma boa representação do tópico de uma consulta, como por exemplo documentos com vizinhos com alta similaridade. Utilizando a sobreposição de agrupamentos de documentos, um documento dominante aparecerá em muitos agrupamentos com uma alta ordenação. Assim como um tópico pode ter muitos subtópicos, o conjunto recuperado pode ser dividido em muitos grupos de subtópicos. Um documento que aparece em todos os subtópicos, provavelmente será subtópico em todos os agrupamentos, assim sendo os autores o chamam de documento dominante. A partir desses documentos dominantes, são selecionados os termos para a expansão que recuperarão documentos relacionados. Assim sendo Lee et al. (2008), selecionam novas amostragem de documentos para a realimentação de relevantes utilizando a técnica de *clustering k-nearest neighbors* (k-NN).

O método de nova amostragem baseado em *clustering* pega os melhores documentos pseudo relevantes é baseado em um modelo de linguagem e no modelo de relevância que mostram ser um caminho útil para se construir um modelo de consulta dos documentos melhores classificados. O ponto essencial desta proposta é que um documento que aparece em múltiplos *clustering* melhores classificados contribui mais para termos da consulta do que outros documentos.

Lee et al. (2008) apresentam os passos para o processo de nova amostragem de documentos da seguinte maneira: documentos são recuperados por uma dada consulta pelo modelo de linguagem probabilística capaz de analisar uma seqüência de palavras gerando partes de um texto. Na recuperação de informação, o modelo de linguagem utiliza documentos como modelos e uma consulta como *string* do texto gerado dos modelos de documentos. O modelo probabilístico de consulta estima modelos de linguagem de documentos utilizando o avaliador máximo probabilístico. Os documentos podem ser ordenados pela geração probabilística de novas amostragens de consultas dos modelos de linguagem de documentos.

O próximo passo segue com a geração dos *clustering* utilizando o método k-NN para a recuperação dos N (100 documentos) documentos para encontrar entre eles os documentos ”dominantes”. Um documento pode pertencer a mais de um *clustering*.

No *clustering k-NN*, cada documento desempenha um papel central no sentido de formar seu próprio *clustering* com seus k vizinhos mais próximos pela similaridades entre eles. Os autores representam um pela pesagem *tfidf* e normalização cosseno. A similaridade cosseno é utilizada para calcular similaridades entre os documentos recuperados melhores classificados.

Lee et al. (2008) têm como hipótese em que um documento dominante pode possuir muitos visinhos com similaridade alta, participando de muitos *clustering*. Por outro lado documentos pertencentes a um único *clustering* podem não ter vizinhos com alta similaridade devido a ruídos como polissemias ou termos genéricos. *Clustering* de documentos também podem refletir a associação de termos e documentos do cálculo da similaridade. Neste trabalho, se um documento pertence a muitos *clustering* e os *clustering* são altamente relacionados com a consulta, os autores assumem isto como sendo um documento dominante. Uma nova amostragem baseada em *clustering* é repetidamente alimentada com documentos dominantes baseados nos

clusters de documentos.

Após a formação dos *clusters*, os autores ordenam os mesmos pelo modelo de linguagem baseado em *cluster*. Os documentos no topo do *ranking* dos *clusters* são utilizados para a realimentação. Note que os *clusters* são utilizados somente para a seleção dos documentos. Finalmente os termos que serão utilizados para a expansão da consulta original são selecionados com base no modelo de relevância para cada documento nos clusters melhor ranqueados. O modelo de relevância é uma distribuição multinominal na qual estima a probabilidade do termo w dado uma consulta Q.

Para avaliar a proposta (Lee et al., 2008) realizaram alguns experimentos utilizando cinco corpus do TREC: (i)*ROBUST*, (ii) *AP*, (iii) *WSJ*, (iv) *GOV2* e (v) *WT10g*. Sendo os três primeiros, corpus de tamanho pequeno (contendo notícias) e os dois últimos são coleções web consideradas grandes. Para medir a eficiência da proposta nos experimentos foi utilizado a medida MAP, apresenta na Seção 2.2.

Ao final do trabalho podemos observar que a utilização de novas amostragem de documentos baseadas em grupos é uma proposta eficiente quando utilizado em coleções grandes, pois Lee et al. (2008) obtiveram nos resultados dos experimentos em coleções com essas características ganho em todos os experimentos realizados. Nas coleções *GOV2* e o ganho foi de 16,82% e 6,28% respectivamente comparado aos resultados do modelo de linguagem e ao modelo de relevância. Na coleção *WT10g* o ganho foi de 16,63% e 26,38% comparados respectivamente com o modelo de linguagem e o modelo de relevância. Já os resultados em coleções pequenas não foram tão bons.

Com o estudo do trabalho apresentado por Lee et al. (2008), ficou claro que a utilização de técnicas de agrupamento para a definição dos documentos a serem utilizados na EC junto ao Modelo TR+ é inviável, uma vez que o corpus de documentos utilizado nesta dissertação é de um tamanho pequeno se comparado com corpus utilizados habitualmente em avaliações de RI, como os presentes por exemplo no *CLEF*. Apesar disso a utilização de técnicas de agrupamentos para a expansão de consulta pode ser visto como um futuro teste a ser realizado desde que possamos trabalhar com um corpus de tamanho maior.

5.6 Considerações sobre o capítulo

No Capítulo 5 apresentamos alguns trabalhos estudados durante o desenvolvimento da dissertação. Os trabalhos apresentados foram selecionados por terem contribuído para a conclusão da dissertação. Os trabalhos apresentados neste nos possibilitou ter um maior conhecimento da aplicação na prática de EC utilizando RR e PRR.

O estudo do trabalho realizado por Custis e Al-Kofahi (2007), identificamos a importância e viabilidade de uso da técnica de EC Pseudo Realimentação de Relevantes para a aplicação em conjunto com o Modelo TR+ na recuperação de informação. Unido a isso, outra importante

contribuição do trabalho apresentado por Custis e Al-Kofahi (2007) foi trazer a oportunidade de um melhor conhecimento de uma situação de uso da fórmula OKAPI.

O trabalho apresentado por Vechtomova e Karamuftuoglu (2007) foi de grande valor para a formulação de nossa proposta, pois ofereceu uma visão prática da utilização da técnica de EC pseudo realimentação de relevantes, chamada neste trabalho por Vechtomova e Karamuftuoglu (2007) de *Blind Feedback*. O trabalho também nos apresentou a utilização de *snippets* dos documentos para a EC, a utilização de *snippets* associada ao Modelo TR+, é uma alternativa à proposta apresentada nesta dissertação. Entretanto devido a necessidade de modificações no Modelo TR+, não lançaremos mão de tal abordagem. O trabalho apresentado por Vechtomova e Karamuftuoglu (2007), fortaleceu a utilização da PRR como uma técnica de EC a ser aplicada junto ao Modelo TR+.

Ao término do estudo do trabalho proposto por Chirita e Nejdl (2007), ficou clara a dificuldade de se aplicar a EC utilizando informações contidas na máquina do usuário ao Modelo TR+, uma vez que, para o uso dessa referência seria fundamental dispor de um *thesaurus* externo como a *WordNet*. Com isso a aplicação dessa proposta ao Modelo TR+ foi descartada pelo tempo exigido para a construção de um *thesaurus* à língua portuguesa. O trabalho proposto por Chirita e Nejdl (2007) fortaleceu a nossa decisão de utilizar a técnica de EC pseudo realimentação de relevantes, eliminando a participação do usuário para melhorar a qualidade das informações recuperadas junto ao Modelo TR+.

Com o estudo do trabalho apresentado por (Orengo & Huyck, 2006) pudemos nos familiarizar com o método de realimentação de relevantes, assim como, com a forma de avaliar a relevância dos documentos recuperados pela consulta original. Assim sendo, pudemos desenvolver os experimentos com RR em conjunto com o Modelo TR+ à RI aplicados nesta dissertação.

Ao estudarmos o trabalho apresentado por Lee et al. (2008), ficou claro que a utilização de técnicas de agrupamento para a definição dos documentos a serem utilizados na EC junto ao Modelo TR+ é inviável, uma vez que o corpus de documentos utilizado nesta dissertação é de um tamanho pequeno se comparado com corpus utilizados habitualmente em avaliações de RI, como os presentes por exemplo no *CLEF*. Apesar disso a utilização de técnicas de agrupamentos para a expansão de consulta pode ser visto como um futuro teste a ser realizado desde que possamos trabalhar com um corpus de tamanho maior.

No capítulo a seguir, Capítulo 6, apresentamos os experimentos realizados para a conclusão do trabalho desenvolvido nesta dissertação. Na Seção 6.1 apresentamos os experimentos realizados por Gonzalez (2005), que foram utilizados por nós no contexto da dissertação como *baseline*. Na Seção 6.5 apresentamos 7 experimentos planejados e realizados para avaliar o desempenho da EC com PRR. Na Seção 6.7 apresentamos 7 experimentos realizados para avaliar o desempenho da EC com RR.

6 Experimentos e Resultados

Nesta seção apresentaremos os experimentos realizados por Gonzalez (Gonzalez, 2005) para o Modelo TR+ utilizados no contexto dessa dissertação como *baseline*, e também os experimentos planejados para avaliar a aplicação de expansão de consulta utilizando as técnicas Pseudo Realimentação de Relevantes e Realimentação de Relevantes em conjunto ao Modelo TR+.

Para a realização dos experimentos utilizamos como ponto de partida os resultados obtidos por Gonzalez em seus experimentos, ou seja, utilizamos as consultas realizadas e os documentos recuperados resultantes destas, e a partir destes resultados aplicamos a expansão de consulta. Para a avaliação dos experimentos foram utilizadas as seguintes métricas apresentadas na Seção 2.2, (i) Precisão, (ii) Abrangência e (iii) MAP. No contexto dessa dissertação, no que tange a avaliação dos experimentos, foram realizados as análises: (i) o número de RLBs utilizado para a EC; (ii) o tipo de RLBs utilizado para a EC, (iii) o número de termos utilizado para a EC.

6.1 Experimento com o Modelo TR+ (Gonzalez, 2005)

Gonzalez em seu experimento para validação do Modelo TR+ no que diz respeito ao contexto da recuperação de informação, empregou a metodologia utilizada nas TRECs (*Text Retrieval Conferences*) (Voorhees, 2005). Foram realizadas 50 consultas referentes a 50 tópicos distintos utilizando o corpus anotado de notícias do Jornal Folha de São Paulo denominado *Folha94*. O corpus *Folha94* consiste de 4.156 artigos do ano de 1994, cada artigo podendo ser classificado com um ou mais assuntos. Cada tópico é representado por um *título*, uma *descrição* e uma *narrativa*, os quais indicam as características que identificam um documento como relevante. O Anexo D, apresenta um exemplo de tópico de consulta utilizado por Gonzalez (2005).

Após a recuperação dos documentos através de diferentes estratégias de RI, Gonzalez utilizou o método de *pooling* (Buckley & Voorhees, 2004), para julgar a relevância dos documentos recuperados. Nesta etapa, os 100 primeiros documentos recuperados para cada tópico, em cada estratégia de RI, foram agrupados e ordenados.

Destes documentos agrupados e ordenados, foram descartados os que não representavam informação relevante, conforme a análise da presença dos termos da consulta no título e corpo do texto. Os documentos restantes foram enviados aos avaliadores em um número de até 10

documentos por avaliador. Finalmente os avaliadores identificaram os documentos relevantes do conjunto de documentos julgados para cada consulta indicada. A avaliação do Modelo TR+ foi finalizada com o cálculo de métricas consolidadas para se mensurar sistemas de recuperação de informação. Como métricas foram utilizados o cálculo da precisão e abrangência, além do cálculo da medida MAP (*Mean Average Precision*)(Buckley & Voorhees, 2004).

Ao analisarmos o experimento realizado por Gonzalez podemos identificar as consultas utilizadas, os termos e as relações lexicais binárias que foram geradas para cada consulta, assim como seus respectivos pesos e os documentos recuperados por elas.

Assim sendo, por exemplo, para a consulta "*abuso sexual*" identificamos os termos e relações lexicais binárias com seus respectivos pesos conforme apresentamos na Figura 11.

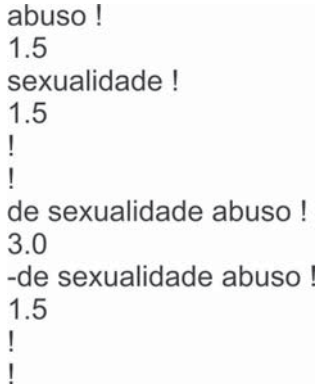

Figura 11 – RLBs e termos e seus pesos gerados para a consulta *"abuso sexual"* pelo Modelo TR+

Com as RLBs e termos da consulta apresentadas na Figura 11, podemos recuperar uma lista de documentos ordenados de acordo com sua relevância. Para cada documento recuperado é definida uma lista de RLBs e termos ordenados. Assim, para a consulta "*abuso sexual*", obtemos a lista de documentos parcialmente apresentada na Tabela 5.

Tabela 5 – Os cinco primeiros documentos recuperados pelo Modelo TR+ com a consulta *"abuso sexual"*

consulta	documento	ordem do documento	peso do documento
301	407	0	1,000000
301	478	1	0,999864
301	437	2	0,995785
301	271	3	0,990838
301	538	4	0,989192

Na Tabela 5 temos:

- *Consulta*: é o identificador da consulta realizada;
- *Doc*: é o identificador do documento recuperado;
- *Ordem doc* é a ordem de recuperação do documento para uma certa consulta;
- *Peso doc* é o grau de relevância do documento para uma certa consulta.

Para cada documento é gerada então uma lista de RLBs e termos. Podemos exemplificar mostrando na Figura 12 parte da lista de RLBs e termos com seus respectivos pesos para o documento identificado como 407 de acordo com a consulta "*abuso sexual*".

```
de(abuso,menina) 29.00
prisao(abuso,balconista) 16.00
prisao.por(balconista,abuso) 16.00
com(sexo,menina) 12.00
por(prisao,abuso) 11.00
de(prisao,balconista) 11.00
```

Figura 12 – RLBs melhor classificadas geradas pelo Modelo TR+ de um documento recuperado com a consulta "*abuso sexual*"

De posse das consultas realizadas por Gonzalez, dos documentos recuperados para cada consulta e das relações lexicais binárias e termos desses documentos, podemos então realizar os experimentos que verificam o desempenho das técnicas de expansão de consulta pseudo realimentação de relevantes e realimentação de relevantes.

6.1.1 Resultados dos experimentos realizados por Gonzalez para validar o Modelo TR+

A Tabela 6 mostra o resultado dos experimentos realizados por Gonzalez (2005) para avaliar a performance da recuperado de informações utilizando o Modelo TR+. Na Figura 13 é possível visualizar a curva da Precisão x Abrangência. Podemos observar tanto na Tabela 6 quanto na Figura 13, o desempenho do Modelo TR+ no que tange à recuperação de informação, alcançando para a medida MAP 85,09%. O valor atingido pelo Modelo TR+ para a medida MAP, pode ser considerado como nosso *baseline*, e é um resultado bastante expressivo. Podemos observar que, ao desconsiderarmos a abrangência dos documentos recuperados (abrangência = 0), a precisão interpolada atingida foi de 97,33%. Já quando a abrangência dos documentos recuperados é o foco da análise, podemos observar que o Modelo TR+ alcança uma precisão de 48,33%. Estes resultados demonstram a eficiência do Modelo TR+ para a recuperação de documentos nos moldes em que foram conduzidos os experimentos.

Tabela 6 – Precisão x Abrangência dos experimentos do Modelo TR+

abrangência	precisão
0	0,9733
0,1	0,9733
0,2	0,9623
0,3	0,9557
0,4	0,9296
0,5	0,9245
0,6	0,8830
0,7	0,8358
0,8	0,7716
0,9	0,6073
1	0,4836
MAP	0,8509

6.2 Processo de Normalização

Todos os experimentos, tanto com PRR como com RR passaram por um processo de normalização dos pesos dos termos e RLBs das consultas expandidas. O processo de normalização tem o intuito de garantir a uniformidade da faixa de valores dos pesos tanto dos termos como das RLBs utilizados na expansão das consultas. O processo de normalização é realizado com base nos pesos dos termos e RLBs das consultas originais.

O processo de normalização dos pesos dos termos e das RLBs no contexto desse trabalho, utiliza o maior peso encontrado entre os termos e as RLBs de cada consulta.

Para melhor entendimento, seguindo os exemplos trabalhados até o momento, utilizaremos as RLBs e os pesos apresentados na Figura 11. Onde o maior peso das RLBs da consulta *"abuso sexual"* é 3.0 para a RLB do tipo classificação "*de(sexualidade,abuso)*". Os termos e RLBs que foram adicionados a esta consulta tiveram seus pesos normalizados para a faixa de valores entre 0 e 3. Na Figura 14 apresentamos os pesos das RLBs apresentadas na Figura 11 com seus pesos normalizados.

6.3 Julgamento de relevância dos documentos recuperados

Após serem realizados os experimentos, o próximo passo a ser cumprido foi a análise dos documentos recuperados com a aplicação da técnica de expansão de consulta e retiradas das RLBs.

Os passos para o julgamento da relevância dos documentos recuperados pelos experimentos com EC apresentados nas subseções 6.5 e 6.7, são apresentados na Figura 15, e seguem o roteiro abaixo:

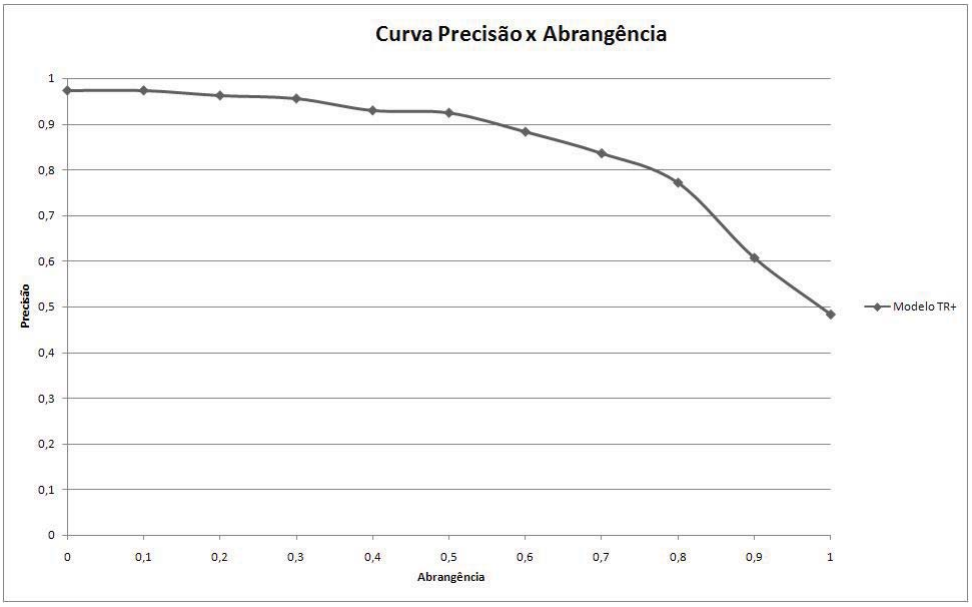

Figura 13 – Curva Precisão x Abrangência para experimento sem EC

```
de(abuso,menina) !
3.00
com(sexo,menina) !
1.2
por(prisao,abuso) !
1.1
de(prisao,balconista) !
1.1
!
prisao(abuso,balconista) !
1.6
prisao.por(balconista,abuso) !
1.6
```

Figura 14 – Pesos normalizados das RLBs apresentados na Figura 11

- (i) O primeiro passo tem início após a recuperação dos documentos para a cada consulta. De posse dos documentos recuperados, é feita a exclusão dos documentos que já tinham sido recuperados e julgados por Gonzalez em seus experimentos. Esse passo é realizado automaticamente por um programa de computador desenvolvido para tal finalidade;

- (ii) O segundo passo após a eliminação dos documentos já julgados por Gonzalez consiste em restringir o total de documentos em um número igual a 100 documentos para cada consulta sendo estes os 100 documentos mais relevantes. Estes 100 documentos passam por um processo de ordenação levando em conta o seu número de identificação;

- (iii) O terceiro passo é a eliminação dos documentos que, dentre estes 100 documentos, não contenham informação relevante à consulta;

- (iv) Os documentos restantes foram divididos em grupos de até 10 documentos, sendo enviados por email a avaliadores para o julgamento da sua relevância;

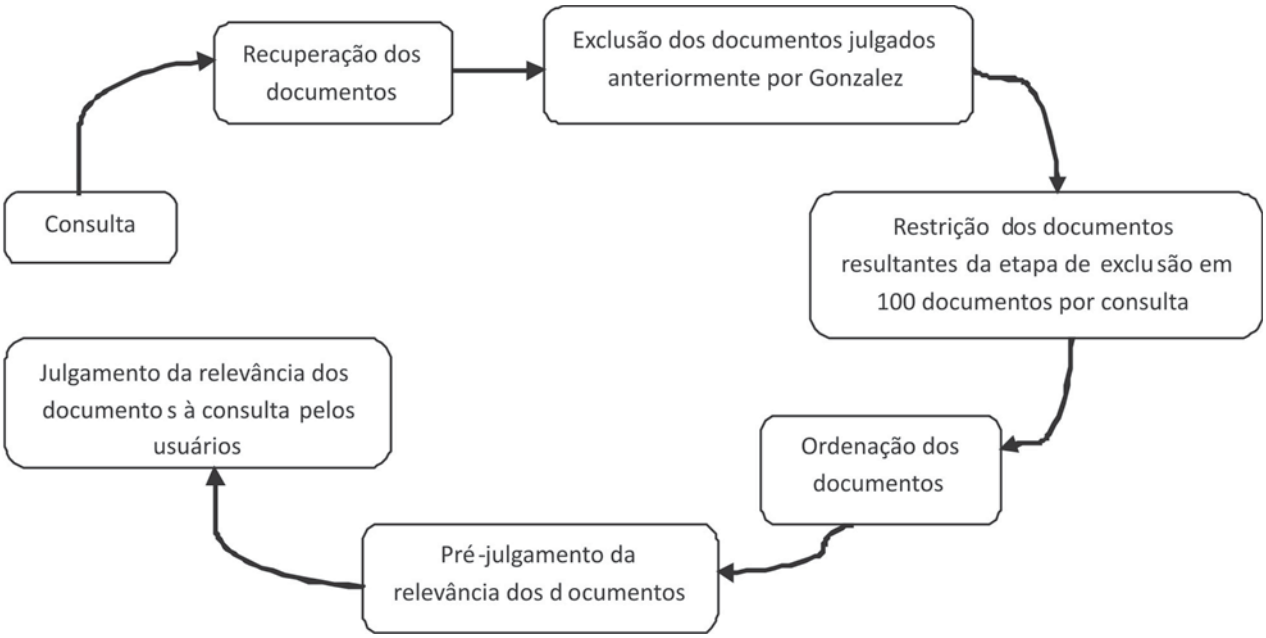

Figura 15 – Processo de julgamento de relevância dos documentos recuperados

- (v) Aos avaliadores cabe a tarefa de analisar o conjunto de documentos que lhes foi enviado, marcando os documentos como:
  a. Relevante
  b. Não Relevante.

Para o avaliador considerar um documento como sendo relevante para uma dada consulta, é necessário que o documento contenha alguma informação digna de ser mencionada em um relatório sobre o tópico da consulta.

Excetuando-se a avaliação da relevância dos documentos, que deve ser realizada de forma manual pelos avaliadores, os demais passos são realizados utilizando apoio automatizado. O passo (i) foi planejado para que não se repetisse o trabalho de julgamento realizado por Gonzalez e assim diminuir o tempo de execução dos experimentos. Os passos (ii) e (iv) foram planejados para desvincular os documentos de qualquer experimento realizado, aumentando assim a isenção nos resultados. O passo (iii) foi planejado para diminuir o número de documentos a serem julgados, e pode ser assim descrito:

1. Os termos da consulta foram previamente colocados em destaque quando estes aparecerem no texto dos documentos;

2. Os documentos são descartados quando, além de não possuirem os termos da consulta em destaque dentro de seu texto, e não apresentarem dúvidas quanto a sua não relevância ao serem lidos, o título, a primeira e última frase do texto;

3. Os documentos devem ser enviados para avaliação quando apresentar os termos da consulta em seu texto, ou ainda quando houver alguma dúvida quanto a sua relevância.

O resultado desses passos foi uma lista de 130 documentos agrupados em 13 conjuntos de 10 documentos cada, independente do tópico da consulta. Cada conjunto de documentos deve ser enviado para o julgamento de um avaliador (veja exemplo do instrumento de avaliação C). De posse dessas avaliações, podemos mensurar a qualidade da recuperação realizada pelo Modelo TR+ com expansão de consulta.

6.4 Etapas da aplicação das técnicas de expansão de consulta PRR e RR em conjunto com o Modelo TR+

Nesta seção apresentaremos passo á passo a aplicação das técnicas de PRR e de RR em conjunto ao Modelo TR+. As técnicas de expansão de consulta já apresentadas nas seções 3.2 e 3.2.4 foram utilizadas seguindo as consultas realizadas por Gonzalez em seus experimentos junto ao Modelo TR+. Tendo as consultas realizadas por Gonzalez e conseqüentemente os documentos recuperados para cada consulta podemos assim realizar os experimentos planejados e apresentados nas seções 6.5 e 6.7.

6.4.1 Passos da aplicação da técnica de expansão de consulta PRR em conjunto ao Modelo TR+

Para a utilização da técnica de EC PRR em conjunto ao Modelo TR+ seguimos os seguintes passos:

1. Definição dos 3 documentos melhor classificados para cada consulta;
2. Extração dos Termos e RLBs de cada consulta de acordo com os critérios planejados para os experimentos apresentados na Seção 6.5;
3. Realimentação das consultas utilizando os Termos e RLBs extraídos dos 3 documentos melhor classificados para cada consulta de acordo com os critérios apresentados na Seção 6.5;
4. Normalização dos pesos dos Termos e RLBs adicionados a cada consulta original;
5. Reformulação das consultas com os novos Termos e RLBs de acordo com os critérios apresentados na Seção 6.5,
6. Eliminação dos documentos já julgados nos experimentos realizado por Gonzalez;
7. Pré-análise dos documentos, eliminando aqueles que não continham os termos da consulta original no seu título e corpo;

8. Análise dos documentos restantes por avaliadores humanos quanto a sua relevância para as consultas correspondentes;

9. Cálculo das métricas de avaliação: (i) precisão, (ii) abrangência e (iii) MAP.

Na Figura 16 podemos observar uma ilustração das etapas da aplicação da técnica de expansão de consulta PRR em conjunto ao Modelo TR+.

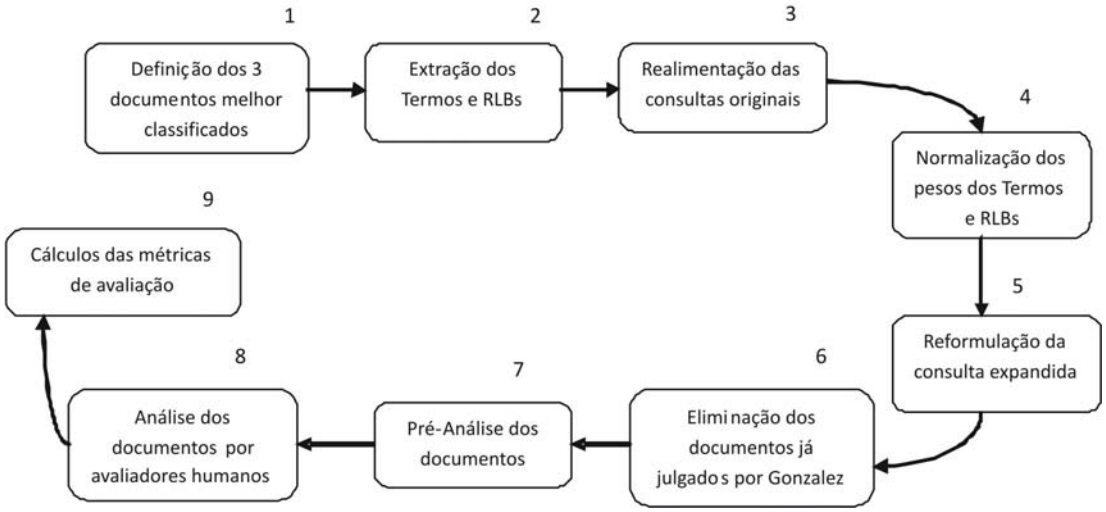

Figura 16 – Etapas da aplicação da técnida de EC PRR em conjunto com o Modelo TR+

6.4.2 Passos da aplicação da técnica de expansão de consulta RR em conjunto com o Modelo TR+

Para a utilização da técnica de EC RR em conjunto com o Modelo TR+ seguimos os seguintes passos:

1. Definição por avaliadores humanos dos 3 documentos mais relevantes de uma lista de 10 documentos à cada consulta;

2. Extração dos Termos e RLBs de cada consulta de acordo com os critérios planejados para os experimentos apresentados na Seção 6.5;

3. Realimentação das consultas utilizando os Termos e RLBs extraídos dos 3 documentos melhor classificados para cada consulta de acordo com os critérios apresentados na Seção 6.5;

4. Normalização dos pesos dos Termos e RLBs adicionados a cada consulta original;

5. Reformulação das consultas com os novos Termos e RLBs de acordo com os critérios apresentados na Seção 6.5,

6. Eliminação dos documentos já julgados nos experimentos realizado por Gonzalez;

7. Pré-análise dos documentos, eliminando aqueles que não continham os termos da consulta original no seu título e corpo;

8. Análise dos documentos restantes por avaliadores humanos quanto a sua relevância para as consultas correspondentes;

9. Cálculo das métricas de avaliação: (i) precisão, (ii) abrangência e (iii) MAP.

Na Figura 17 podemos observar uma ilustração das etapas da aplicação da técnica de expansão de consulta RR em conjunto ao Modelo TR+.

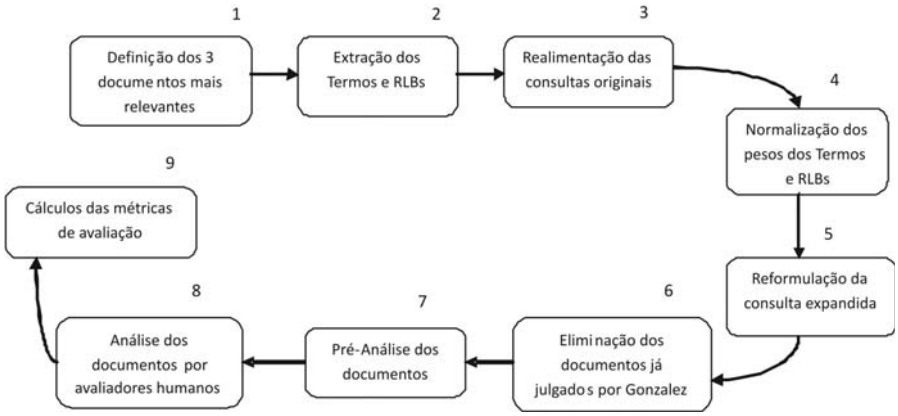

Figura 17 – Etapas da aplicação da técnida de EC RR em conjunto com o Modelo TR+

6.5 Experimentos com o Modelo TR+ Utilizando Pseudo Realimentação de Relevantes

Os experimentos para análise da aplicação da expansão de consulta ao processo descrito assumem que as n RLBs e os m termos dos três documentos recuperados são informações relevantes de acordo com a consulta original. Com isso podemos expandir a consulta original utilizando um critério pré-definido. Os experimentos realizados com PRR serão apresentados nas subseções 6.5.1, 6.5.2, 6.5.3, 6.5.4, 6.5.5 e 6.5.6. Em todos os experimentos foi realizado um processo de normalização dos pesos das RLBs e termos utilizados para expandir a consulta original, apresentado em maiores detalhes na Seção 6.2. Na Figura 18 apresentamos o processo de expansão de consulta utilizado para todos os experimentos.

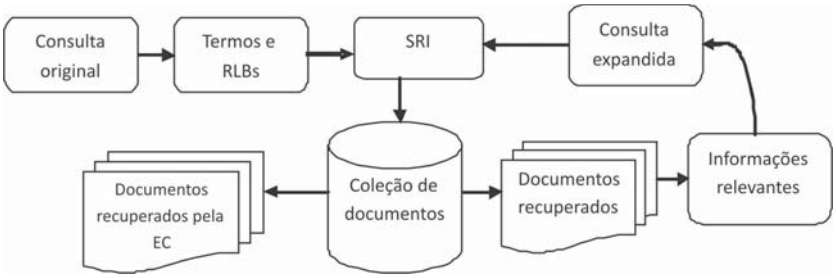

Figura 18 – Processo de expansão de consulta utilizado nos experimentos

6.5.1 Experimento 1 com PRR

Descrição: No primeiro experimento, expandimos as consultas com as três RLBs mais ”pesadas” (RLBs com maior evidência nos documentos) provenientes dos três primeiros documentos recuperados, independentemente do tipo dessas, sejam elas Restrições, Associações ou Classificações.

Objetivo: O objetivo desse experimento é analisar se há aumento do desempenho da recuperação dos documentos quando a consulta é expandida utilizando para isso RLBs.

6.5.2 Experimento 2 com PRR

Descrição: No segundo experimento, expandimos as consultas utilizando as RLBs quanto ao seu tipo. Para este experimento utilizamos os três documentos melhor classificados oriundos da consulta original. A perspectiva desse experimento é identificar se diferentes tipos de RLBs podem obter resultados melhores que outros. Assim após a consulta original, identificam-se os três documentos mais pesados e as três RLBs de acordo com o seu tipo: *Restrição*, *Associação* ou *Classificação*. Novamente a consulta original é realimentada com as RLBs escolhidas. Este experimento se subdivide em três outros, que diferem como segue:

1. Expansão da consulta original utilizando as três RLBs mais pesadas do tipo *Restrição* dos três documentos melhor classificados oriundos da consulta original.

2. Expansão da consulta original utilizando as três RLBs mais pesadas do tipo *Classificação* dos três documentos melhor classificados oriundos da consulta original.

3. Expansão da consulta original utilizando as três RLBs mais pesadas do tipo *Associação* dos três documentos melhor classificados oriundos da consulta original.

Objetivo: O objetivo desse experimento é verificar se o tipo de RLB utilizada para a expansão da consulta original possui influência na qualidade da recuperação.

6.5.3 Experimento 3 com PRR

Descrição: O terceiro experimento planejado para avaliar o desempenho do processo de expansão de consulta realimenta a consulta original com os três termos mais ”pesados” dos três documentos melhor classificados oriundos da consulta original.

A realimentação da consulta original utilizando os termos mais pesados segue a rotina especificada nos experimentos anteriores, ou seja, após a consulta original são identificados os três documentos mais relevantes e assim os três termos mais pesados desses documentos são utilizados pra realimentar a consulta original.

Objetivo: Com esse experimento temos como objetivo avaliar o desempenho da expansão da consulta utilizando para isso somente termos resultantes da primeira etapa do processo.

6.5.4 Experimento 4 com PRR

Descrição: O Experimento 4 avalia a utilização das cinco RLBs mais ”pesadas” provenientes dos três documentos recuperados melhor classificados, oriundos da consulta original. As RLBs utilizadas neste experimento independen do seu tipo, sejam elas Restrições, Associações ou Classificações.

Objetivo: O objetivo desse experimento é avaliar se o desempenho da recuperação dos documentos é influenciada, quando a consulta é expandida utilizando para isso um número maior de RLBs relevantes à necessidade do usuário.

6.5.5 Experimento 5 com PRR

Descrição: Experimento 5 avalia a utilização dos cinco termos mais ”pesados” provenientes dos três documentos recuperados melhor classificados oriundos da consulta original.

Objetivo:O objetivo desse experimento é avaliar se o desempenho da recuperação dos documentos é aumentado, quando a consulta é expandida utilizando para isso um número maior de termos em comparação ao número de termos utilizados no Experimento 3.

6.5.6 Experimento 6 com PRR

Descrição: O Experimento 6 avalia utilização das dez RLBs mais ”pesadas” provenientes dos três documentos recuperados melhor classificados oriundos da consulta original. As RLBs utilizadas neste experimento assim como nos experimentos 1 e 4 independem do seu tipo, sejam

elas Restrições, Associações ou Classificações.

Objetivo: Esse experimento tem como objetivo avaliar se existe aumento no desempenho de documentos recuperados quando a consulta é expandida utilizando para isso um número maior de RLBs relevante à necessidade do usuário quando comparado com o número de RLBs utilizadas nos experimentos 6.5.1 e 6.5.4.

6.6 Resultados dos Experimentos realizados junto ao Modelo TR+ utilizando Pseudo Realimentação de Relevantes

Nesta seção apresentaremos os resultados dos experimentos realizados junto ao Modelo TR+, adicionando RLBs e termos as consultas originais, utilizando para isso a técnica de expansão de consultas Pseudo Realimentação de Relevantes. Na Tabela 7 e na Figura 19 expomos os resultados obtidos pelos experimentos que adicionaram RLBs as consultas originais em comparação aos resultados do *baseline* (Gonzalez, 2005). Já na Tabela 8 e na Figura 20 apresentamos os resultados da expansão das consultas utilizando termos em comparação aos resultados obtidos pelo *baseline*.

Tabela 7 – Resultados dos experimentos adicionando RLBs com PRR

	TR+	Exp 1	Exp 2.1	Exp 2.2	Exp 2.3	Exp 4	Exp 6
Abr	Pr	Pr	Pr	Pr	Pr	Pr	Pr
0	0,9733	0,9733	0,9733	0,9733	0,9733	0,7933	0,9733
0,1	0,9733	0,9725	0,9725	0,9725	0,9725	0,7925	0,9725
0,2	0,9623	0,9561	0,9561	0,9561	0,9561	0,7761	0,9561
0,3	0,6557	0,9401	0,9401	0,9401	0,9401	0,7601	0,9401
0,4	0,9296	0,8946	0,8946	0,8946	0,8946	0,7196	0,8946
0,5	0,9245	0,8748	0,8748	0,8748	0,8748	0,7040	0,8748
0,6	0,8830	0,8049	0,8049	0,8049	0,8049	0,6559	0,8049
0,7	0,8358	0,7654	0,7654	0,7654	0,7654	0,6196	0,7654
0,8	0,7716	0,7066	0,7066	0,7066	0,7066	0,5657	0,7066
0,9	0,6073	0,5464	0,5464	0,5464	0,5464	0,4332	0,5464
1	0,4836	0,3924	0,3924	0,3924	0,3924	0,3200	0,3924
MAP	0,8509	0,8087	0,8087	0,8087	0,8087	0,6526	0,8087

6.6.1 Resultados do Experimento 1 com PRR

Na Tabela 7 e na Figura 19, podemos observar que os resultados do Experimento 1, mantêm o mesmo comportamento dos resultados de Gonzalez (2005) ao compararmos a curva entre a precisão e abrangência. Entretanto para a medida MAP, o Experimento 1 foi de 80,87%, mais

Tabela 8 – Resultados dos experimentos adicionando Termos com PRR

	TR+	Exp 3	Exp 5
Abr	Pr	Pr	Pr
0	0,9733	0,2975	0,2270
0,1	0,9733	0,2697	0,1877
0,2	0,9623	0,2486	0,1866
0,3	0,6557	0,2476	0,1866
0,4	0,9296	0,2399	0,1798
0,5	0,9245	0,2375	0,1772
0,6	0,8830	0,2299	0,1752
0,7	0,8358	0,2060	0,1688
0,8	0,7716	0,1779	0,1598
0,9	0,6073	0,0856	0,0680
1	0,4836	0,0510	0,0394
MAP	0,8509	0,1693	0,1132

de 4 pontos percentuais inferior ao *baseline*. Uma explicação para tal comportamento é o fato que RLBs são particulares de certos documentos (90% das RLBs estão presentes somente em um único documento), e uma vez que as RLBs são retiradas de documentos oriundos da consulta que será expandida, estas RLBs só fortalecem a recuperação dos mesmos documentos. Outra explicação para a media MAP neste experimento ser inferior ao *baseline* é que ao utilizarmos para a EC RLBs como por exemplo, ”*prisao.por(balconista,abuso)*”, o SRI recuperará documentos que não são relevantes a consulta original ”abuso sexual”. Neste exemplo, muito embora a RLB utilizada possua em seus argumentos o termo ”abuso”, esta RLB se refere na realidade ”a prisão do balconista por algum tipo de abuso”.

Reforçando esta conclusão está a análise realizada ao estudarmos as consultas expandidas. Ao término da análise das RLBs utilizadas para a expansão das consultas do Experimento 1 concluímos que: (i) foram utilizadas 450 RLBs nas 50 consultas expandidas; (ii) das 450 RLBs utilizadas somente 94 (20,88%) foram consideradas relevantes ao tópico da consulta; (iii) das 450 RLBs utilizadas para a expansão das consultas, 356 RLBs (79,12%) foram consideradas irrelevantes para o tópico da consulta. Estes números apontam a utilização de poucas RLBs relevantes para a expansão das consultas selecionadas para este experimento.

6.6.2 Resultados dos Experimentos 2.1, 2.2 e 2.3 com PRR

Na Tabela 7 e na Figura 19 podemos observar que os resultados obtidos pelo Experimento 2 em suas variações alcançaram o mesmo desempenho do Experimento 1 (Seção 6.5.1), ou seja, MAP igual a 80,87%. Podemos observar que a utilização de PRR com os diferentes tipos de RLBs para a expansão da consulta, no que tange aos experimentos realizados, não tiveram impacto ao resultado obtido pelo Experimento 1. Uma explicação para tal comportamento é

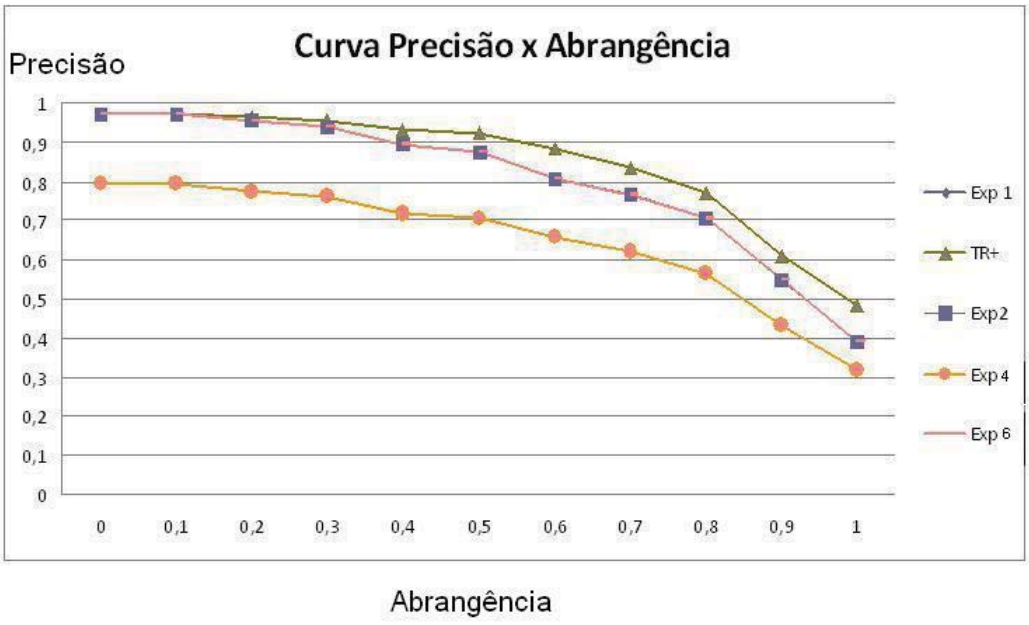

Figura 19 – Curva Precisão x Abrangência para os experimentos que utilizaram RLBs para a EC em conjunto ao Modelo TR+ com PRR

que, ao expandirmos as consultas utilizando os diferentes tipos de RLBs (Restrição, Associação e Classificação), as RLBs já existentes na consulta original não foram retiradas e com isso estas RLBs puderam exercer influência no resultado da recuperação.

Além disso, após a análise das RLBs utilizadas para a expansão das consultas nos experimentos podemos concluir que: (i) no Experimento 2.1 foram utilizadas 450 RLBs, sendo 87 RLBs (19,33%) consideradas relevantes aos seus respectivos tópicos de consulta e 363 RLBs (80,66%) não relevantes; (ii) no Experimento 2.2 foram utilizadas 450 RLBs para a expansão das consultas, sendo 72 RLBs (16%) consideradas relevantes aos seus respectivos tópicos de consulta e 378 RLBs (84%) não relevantes; (iii) no Experimento 2.3 foram utilizadas 450 RLBs para a expansão das consultas, sendo 65 RLBs (14,44%) consideradas relevantes aos seus respectivos tópicos de consulta e 385 RLBs (85,55%) não relevantes. Esta análise aponta a utilização de poucas RLBs relevantes para a expansão das consultas selecionadas para este experimento.

6.6.3 Resultados do Experimento 3 com PRR

Na Tabela 8 e na Figura 20, podemos observar que o comportamento do Experimento 3, que foi bastante diferente dos resultados obtidos tanto por Gonzalez como no Experimento 1 e 2 (utilizando RLBs) . Neste experimento, conforme já exposto, foram utilizados os três

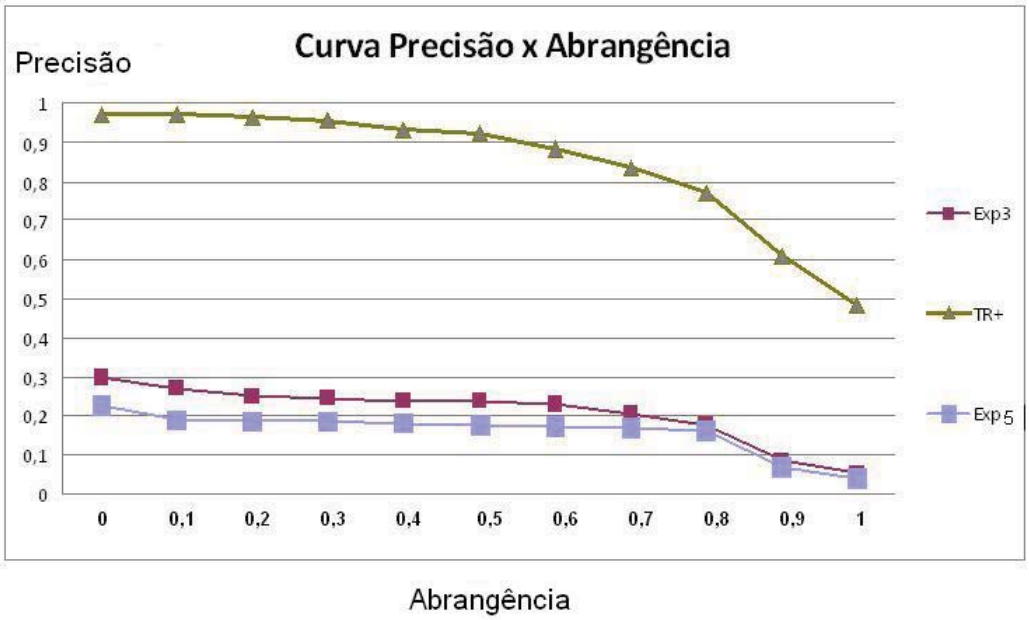

Figura 20 – Curva Precisão x Abrangência para os experimentos que utilizaram Termos para a EC em conjunto ao Modelo TR+ com PRR

termos mais ”pesados” dos três documentos mais relevantes à consulta original. Os resultados apontam uma queda acentuada da precisão nos documentos recuperados em comparação com os experimentos realizados por Gonzalez, alcançando para MAP um valor de 16,93%. Uma explicação para tal comportamento pode ser o fato de que termos são muito genéricos e estão presentes em muitos documentos irrelevantes, e a sua utilização no contexto desse trabalho para RI resultou na recuperação de um número maior de documentos irrelevantes para a consulta em comparação aos Experimentos 1 e 2. Aliado a isso podemos verificar ao analisarmos consulta a consulta os termos utilizados para sua expansão, podemos definir que: (i) foram utilizados 450 termos; (ii) destes somente 78 (17,33%) termos foram considerados relevantes aos tópicos da consulta; (iii) 373 (82,66%) termos foram considerados não relevantes. Essa análise nos leva a crer que aliado ao fato que termos são muito genérico está a utilização para a expansão das consultas de um número muito baixo de termos relevantes, o que leva a um decréscimo considerado para a recuperação de informação.

6.6.4 Resultados do Experimento 4 com PRR

Podemos observar na Tabela 7 e na Figura 19 que o aumento do número de RLBs utilizadas para a expansão da consulta não acarreta em um desempenho melhor para recuperação de in-

formação, sendo seu desempenho de fato prejudicado com este aumento no número de RLBs. O desempenho na recuperação de informação com 5 RLBs dos três documentos melhor classificados na recuperação realizada pela consulta original atingiu para a medida MAP 65,26% contra 80,87% para a mesma medida quando utilizado as 3 RLBs dos três documentos melhor classificados pela recuperação da consulta inicial. Ao analisarmos a relevância das RLBs utilizadas para a expansão das consultas podemos constatar que: (i) foram utilizadas 750 RLBs no total; (ii) destas, 125 (16,66%) foram reconhecidas como relevantes; (iii) do restante das RLBs, ou seja, 625 (83,34%) foram consideradas irrelevantes em comparação aos tópicos das consultas. Ao compararmos o percentual de RLBs relevantes utilizadas para a expansão das consultas deste experimento com o Experimento 1, podemos ver que mesmo com o aumento do número de RLBs o seu percentual foi 4,22% menor. Isto explicaria o fato de que com o aumento do número de RLBs em relação ao Experimento 1, o seu desempenho foi inferior.

6.6.5 Resultados do Experimento 5 com PRR

Os resultados obtidos por este experimento e apresentados na Tabela 8 e na Figura 20, não deixa dúvidas quanto a ineficácia do aumento dos termos de três, no Experimento 3, para cinco termos como proposto nesta rodada de experimentos. Utilizando cinco termos o experimento obteve uma medida MAP de 11,32% diminuindo ainda mais o desempenho já pouco significativo alcançado no Experimento 3 que, para a mesma medida alcançou 16,93%. Uma explicação para tal comportamento, apresentado também no Experimento 3 , pode ser o fato de que termos são muito genéricos e estão presentes em muitos documentos irrelevantes, e a sua utilização no contexto desse trabalho para RI resultou na recuperação de um número maior de documentos irrelevantes para a consulta em comparação aos experimentos 1 e 2 respectivamente. Ao analisarmos consulta a consulta os termos utilizados na sua expansão podemos constatar que: (i) foram utilizados 705 termos para a expansão das consultas; (ii) do total 109 termos foram considerados relevantes; e (iii) 641 termos irrelevantes.

6.6.6 Resultados do Experimento 6 com PRR

Na Tabela 7 e na Figura 19 é possível observar que o aumento do número de RLBs não resulta na melhora do desempenho da recuperação de informação quando comparado com os demais experimentos que utilizaram RLBs apresentados nesta seção. De fato ao se utilizar as 10 RLbs mais pesadas dos três documentos melhor classificados pela recuperação referente à consulta original, este obteve um valor para a medida MAP de 80,87%, sendo este o valor idêntico ao alcançado pelo Experimento 1 onde foram utilizadas as três RLBs para os mesmos três documentos melhor classificados para a consulta original. Aliado a isso, o fato de que das

1500 RLBs utilizadas, somente 265 (17,66%) foram consideradas relevantes e 1235 (82,34%) não relevantes. Esta avaliação nos mostra que o percentual de RLBs relevantes utilizadas para a expansão das consultas é muito parecido ao percentual de RLBs utilizadas no Experimento 1 (20,88%), mesmo que em um número muito maior. Assim confirmamos que o aumento do número de RLBs na expansão das consultas não determina um aumento no desempenho da recuperação de informação se estas RLBs não forem em um número expressivo de RLBs relevantes para as consultas.

6.7 Experimentos realizados junto ao Modelo TR+ utilizando Realimentação de Relevantes

Os experimentos utilizando realimentação de relevantes junto ao Modelo TR+, lançou mão da mesma metodologia dos experimentos realizados por Gonzalez (Seção 6.1) e dos experimentos com pseudo realimentação de relevantes (Seção 6.5), difireciando-se apenas quanto à seleção dos documentos que participaram do processo de EC.

Netes experimentos assumimos que as *n* RLBs e os *m* termos dos três primeiros documentos recuperados (de uma lista de 10 documentos) e julgados como relevantes à consulta original pelos usuários são utilizados para a EC. Com isso podemos expandir a consulta original utilizando um critério pré-definido. Apresentamos os experimentos realizados nas subseções 6.7.1, 6.7.2, 6.7.3, 6.9, 6.7.4, 6.7.5 6.7.6. Todos os experimentos, excetuando-se o experimento 4, apresentado na Subseção 6.9, passaram por um processo de normalização dos pesos das RLBs e termos utilizados no processo de EC. O processo de normalização será apresentado em maiores detalhes na Subseção 6.2.

6.7.1 Experimento 1 com RR

Descrição: No primeiro experimento, utilizamos para a avaliar o processo de EC as três RLBs mais ”pesadas”(RLBs com maior freqüência no documento) provenientes dos três primeiros documentos recuperados e julgados relevantes à consulta original pelo usuário, independente do seu tipo, seja ele Restrição, Associação ou Classificação.

Objetivo: O objetivo desse experimento é avaliar se há impacto no desempenho da recuperação dos documentos quando a consulta é expandida utilizando para isso RLBs relevantes à necessidade do usuário.

6.7.2 Experimento 2 com RR

Descrição: O Experimento 2 avalia a aplicação da expansão de consulta utilizando RLBs quanto ao seu tipo. Para este experimento foram utilizados os três primeiros documentos julgados como relevantes pelo usuário à consulta original. A perspectiva desse experimento é identificar se diferentes tipos de RLBs podem obter resultados melhores que outros. Assim após a consulta original, identificou-se os três documentos mais pesados extraindo as três RLBs de acordo com o seu tipo: *Restrição*, *Associação* ou *Classificação*. Novamente a consulta original é realimentada com as RLBs selecionadas. Este experimento se subdivide em três outros experimentos que diferem como segue:

1. Expande a consulta original utilizando as três RLBs mais pesadas do tipo *Restrição* para os três documentos mais relevantes à consulta.

2. Expande a consulta original utilizando as três RLBs mais pesadas do tipo *Classificação* para os três documentos mais relevantes à consulta.

3. Expande a consulta original utilizando as três RLBs mais pesadas do tipo *Associação* para os três documentos mais relevantes à consulta.

Objetivo: O objetivo desse experimento é verificar se o tipo de RLB utilizada para a expansão da consulta original possui influência na recuperação dos documentos.

6.7.3 Experimento 3 com RR

Descrição: O terceiro experimento avalia o desempenho do processo de expansão de consulta utilizando os três termos mais ”pesados” dos três primeiros documentos julgados como relevantes à consulta original.

A realimentação da consulta original utilizando os termos mais pesados segue a rotina especificada nos experimentos anteriores, ou seja, após a consulta original são identificados os três documentos mais relevantes e assim os três termos mais pesados desses documentos são utilizados pra realimentar a consulta original.

Objetivo: Com esse experimento temos como objetivo avaliar o desempenho da expansão da consulta utilizando para isso somente termos resultantes da primeira etapa do processo.

6.7.4 Experimento 4 com RR

Descrição: O Experimento 4 avalia a utilização das cinco RLBs mais ”pesadas” provenientes dos três primeiros documentos recuperados julgados relevantes à consulta original, inde-

pendente do seu tipo, seja ele Restrição, Associação ou Classificação.

Objetivo: O objetivo desse experimento é avaliar se há aumento do desempenho da recuperação dos documentos quando a consulta é expandida utilizando para isso um número maior RLBs relevantes à necessidade do usuário.

6.7.5 Experimento 5 com RR

Descrição: Experimento 5 avalia a utilização dos cinco termos mais ”pesados” provenientes dos três primeiros documentos recuperados julgados relevantes à consulta original, independente do seu tipo, seja ele Restrição, Associação ou Classificação.

Objetivo: O objetivo desse experimento é analisar se há aumento no desempenho da recuperação dos documentos, quando a consulta é expandida utilizando para isso um número maior de termos relevantes à necessidade do usuário comparados ao número de termos utilizados no experimento 6.7.3.

6.7.6 Experimento 6 com RR

Descrição: O Experimento 6 avalia a utilização das dez RLBs mais ”pesadas” provenientes dos três primeiros documentos recuperados julgados relevantes à consulta original, independente do seu tipo, seja ele Restrição, Associação ou Classificação.

Objetivo: Esse experimento tem como objetivo analisar se existe aumento no desempenho da recuperação quando a consulta é expandida utilizando para isso um número maior de RLBs relevante à necessidade do usuário quando comparado com o número de RLBs utilizadas nos experimentos 6.7.1 e 6.7.4.

6.8 Resultados dos Experimentos realizados junto ao Modelo TR+ utilizando Realimentação de Relevantes

Nesta seção apresentaremos os resultados dos experimentos realizados junto ao Modelo TR+, adicionando RLBs e termos as consultas originais, utilizando para isso a técnica de expansão de consultas Realimentação de Relevantes. Na Tabela 9 e na Figura 21 expomos os resultados obtidos pelos experimentos que adicionaram RLBs as consultas originais em comparação aos resultados do *baseline* (Gonzalez, 2005). Já na Tabela 10 e na Figura 22 apresentamos os resultados da expansão das consultas utilizando termos em comparação aos resultados obtidos pelo *baseline*.

Tabela 9 – Resultados dos experimentos adicionando RLBs com RR

	TR+	Exp 1	Exp 2.1	Exp 2.2	Exp 2.3	Exp 4	Exp 6
Abr	Pr	Pr	Pr	Pr	Pr	Pr	Pr
0	0,9733	0,9733	0,9733	0,9733	0,9733	0,9538	0,9733
0,1	0,9733	0,9725	0,9725	0,9725	0,9725	0,9525	0,9725
0,2	0,9623	0,9561	0,9561	0,9561	0,9561	0,9361	0,9561
0,3	0,6557	0,9401	0,9401	0,9401	0,9401	0,9287	0,9401
0,4	0,9296	0,8946	0,8946	0,8946	0,8946	0,8746	0,8946
0,5	0,9245	0,8748	0,8748	0,8748	0,8748	0,8548	0,8748
0,6	0,8830	0,8049	0,8049	0,8049	0,8049	0,7846	0,8049
0,7	0,8358	0,7654	0,7654	0,7654	0,7654	0,7454	0,7654
0,8	0,7716	0,7066	0,7066	0,7066	0,7066	0,6866	0,7066
0,9	0,6073	0,5464	0,5464	0,5464	0,5464	0,5264	0,5464
1	0,4836	0,3924	0,3924	0,3924	0,3924	0,3924	0,3924
MAP	0,8509	0,8087	0,8087	0,8087	0,8087	0,7901	0,8087

Tabela 10 – Resultados dos experimentos adicionando Termos com RR

	TR+	Exp 3	Exp 5
Abr	Pr	Pr	Pr
0	0,9733	0,2947	0,2170
0,1	0,9733	0,2669	0,1846
0,2	0,9623	0,2452	0,1846
0,3	0,6557	0,2439	0,1846
0,4	0,9296	0,2365	0,1778
0,5	0,9245	0,2341	0,1752
0,6	0,8830	0,2265	0,1740
0,7	0,8358	0,2026	0,1594
0,8	0,7716	0,1745	0,1504
0,9	0,6073	0,0822	0,0680
1	0,4836	0,0476	0,0394
MAP	0,8509	0,1674	0,1101

6.8.1 Resultados do Experimento 1 com RR

Na Tabela 9 e na Figura 21, podemos observar que os resultados do Experimento 1 com RR, mantêm o mesmo comportamento dos resultados de Gonzalez (2005) ao compararmos a curva entre a precisão e abrangência. Também podemos observa que seu comportamento é semelhante ao encontrando no Experimento 1 com PRR com uma medida MAP de 80,87%, mais de 4 pontos percentuais inferior ao *baseline*. A explicação apresentada para esta rodada de experimento com PRR pode ser utilizada neste experimento, uma vez que RLBs são particulares de certos documentos (90% das RLBs estão presentes somente em um único documento), e sendo que as RLBs são retiradas de documentos oriundos da consulta que será expandida, estas RLBs só fortalecem a recuperação dos mesmos documentos. Outra explicação para a media MAP ser

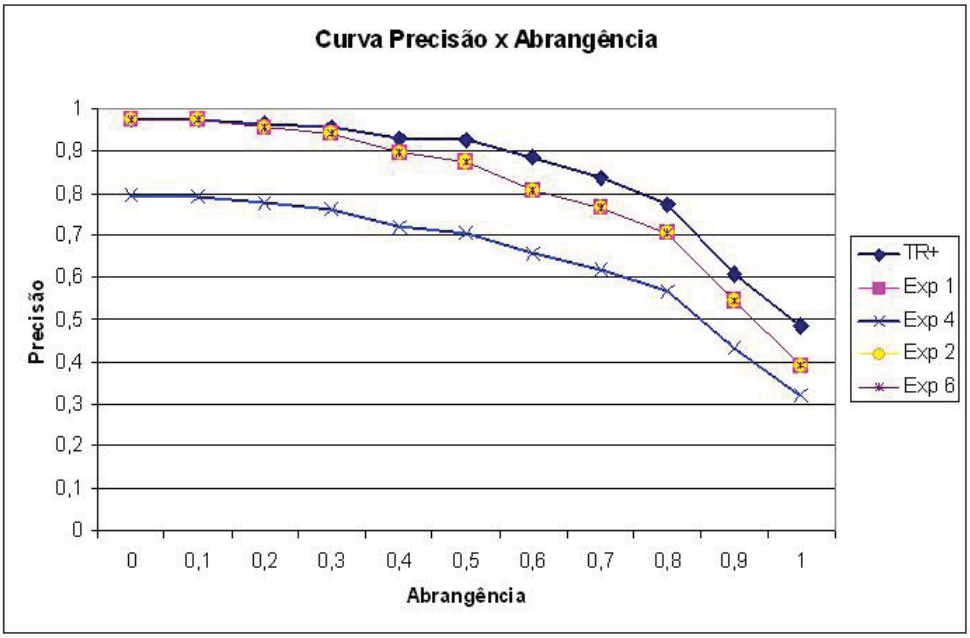

Figura 21 – Curva Precisão x Abrangência para os experimentos que utilizaram RLBs para a EC em conjunto ao Modelo TR+ com RR

inferior ao *baseline* neste experimento apresentado no Experimento 1 com PRR, é que ao utilizarmos para a EC RLBs como por exemplo, "*prisao.por(balconista,abuso)*", o SRI recuperará documentos que não são relevantes a consulta original "abuso sexual". Neste exemplo, muito embora a RLB utilizada possua em seus argumentos o termo "abuso", esta RLB se refere na realidade "a prisão do balconista por algum tipo de abuso".

Reforçando esta conclusão está a análise realizada ao estudarmos as consultas expandidas. Ao término da análise das RLBs utilizadas para a expansão das consultas do Experimento 1 concluímos que: (i) foram utilizadas 450 RLBs nas 50 consultas expandidas; (ii) das 450 RLBs utilizadas somente 91 (20,22%) foram consideradas relevantes ao tópico da consulta; (iii) das 450 RLBs utilizadas para a expansão das consultas, 359 RLBs (79,77%) foram consideradas irrelevantes para o tópico da consulta. Estes números apontam a utilização de poucas RLBs relevantes para a expansão das consultas selecionadas para este experimento.

6.8.2 Resultados dos Experimentos 2.1, 2.2 e 2.3 com RR

Podemos observar que os resultados obtidos pelo Experimento 2 em suas variações alcançaram o mesmo desempenho do Experimento 1 tanto com PRR como com RR (subseções 6.5.1 e 6.7.1), ou seja MAP igual a 80,87%. Como já exposto anteriormente, os diferentes tipos de RLBs na expansão da consulta, no que tange aos experimentos realizados no contexto dessa dissertação, não tiveram impacto quanto ao resultado obtido pelo experimento. Uma explicação para tal comportamento, é que ao expandirmos as consultas utilizando os diferentes tipos

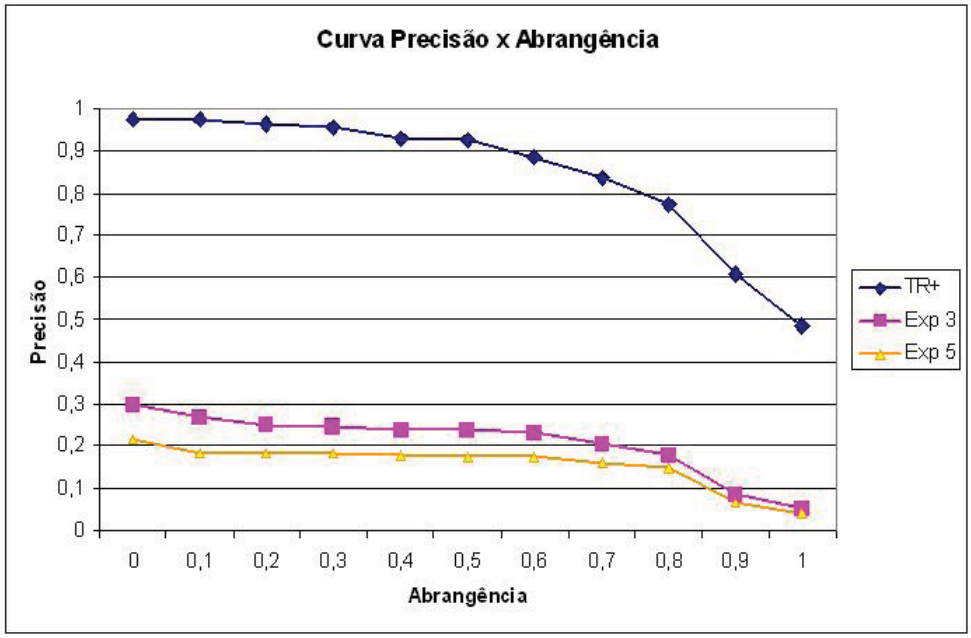

Figura 22 – Curva Precisão x Abrangência para os experimentos que utilizaram Termos para a EC em conjunto ao Modelo TR+ com RR

de RLBs (Restrição, Associação e Classificação), as RLBs já existentes na consulta original não foram retiradas e com isso estas RLBs puderam exercer influência no resultado dos experimentos. Agregado a esta explicação ao fato de que houve uma variação muito pequena de documentos de onde foram extraídas as RLBs que fizeram parte da EC nos experimentos com PRR e RR. Além disso, após a análise das RLBs utilizadas para a expansão das consultas nos experimentos podemos concluir que: (i) no Experimento 2.1 foram utilizadas 450 RLBs, sendo 89 RLBs (19,77%) consideradas relevantes aos seus respectivos tópicos de consulta e 361 RLBs (80,33%) não relevantes; (ii) no Experimento 2.2 foram utilizadas 450 RLBs para a expansão das consultas, sendo 69 RLBs (15,33%) consideradas relevantes aos seus respectivos tópicos de consulta e 381 RLBs (84,77%) não relevantes; (iii) no Experimento 2.3 foram utilizadas 450 RLBs para a expansão das consultas, sendo 60 RLBs (13,33%) consideradas relevantes aos seus respectivos tópicos de consulta e 380 RLBs (86,66%) não relevantes. Esta análise aponta a utilização de poucas RLBs relevantes para a expansão das consultas selecionadas para este experimento.

6.8.3 Resultados do Experimento 3 com RR

Podemos observar nos resultados obtidos no Experimento 3 com RR, que o comportamento desse experimento foi semelhante ao demonstrado pelo Experimento 3 com PRR (Seção 6.5.3), e portanto, bastante diferente dos resultados apresentados por Gonzalez. A utilização dos três termos mais ”pesados” dos três documentos mais relevantes à consulta original com RR, resul-

tou em uma medida MAP com valor de 16,74% contra 16,93% obtidos pelo experimento 3 com PRR. Uma possível explicação para este pequeno decréscimo na medida MAP, se dá pelo fato de que em 8 consultas foram adicionados termos de diferentes documentos aos utilizados para o mesmo experimento com PRR. Tal fato em conjunto com a generalidade dos termos fez que com o resultado do experimento tivesse sido inferior ao executado com PRR. De fato podemos verificar ao analisarmos consulta a consulta os termos utilizados para sua expansão, podemos definir que: (i) foram utilizados 450 termos; (ii) destes somente 81 (18%) termos foram considerados relevantes aos tópicos da consulta; (iii) 369 (82%) termos foram considerados não relevantes. Essa análise nos leva a crer que aliado ao fato que termos são muito genéricos está a utilização para a expansão das consultas de um número muito baixo de termos relevantes, o que leva a um decréscimo considerado para a recuperação de informação.

6.8.4 Resultados do Experimento 4 com RR

Podemos observar com este experimento, que o aumento do número de RLBs utilizadas para a expansão da consulta não acarreta em um desempenho melhor para recuperação dos documentos, ao ser comparado com os resultados obtidos por Gonzalez. Entretanto, o desempenho na recuperação dos documentos com 5 RLBs dos três documentos melhor classificados na recuperação realizada pela consulta original com RR atingiu para a medida MAP 79,01% contra 65,26% para a mesma medida, quando utilizado para o mesmo experimento PRR. A utilização de RR com 5 RLBs teve um ganho de 13,75%, isto nos leva a crer que os novos documentos utilizados no experimento pela RR influenciou positivamente o seu resultado, melhorando seu desempenho na recuperação. Ao analisarmos a relevância das RLBs utilizadas para a expansão das consultas podemos constatar que: (i) foram utilizadas 750 RLBs no total; (ii) destas, 142 (18,93%) foram reconhecidas como relevantes; (iii) do restante das RLBs, ou seja, 608 (81,06%) foram consideradas irrelevantes em comparação aos tópicos das consultas. Podemos observar que o aumento no número de RLBs relevantes utilizadas neste experimento em comparação ao número de RLBs utilizadas no Experimento 4 com PRR resultou no aumento da performance no que tange a medida MAP.

6.8.5 Resultados do Experimento 5 com RR

Os resultados obtidos por este experimento e apresentados na Tabela 10 e na Figura 22, não deixa dúvidas quanto a ineficácia do aumento dos termos de três, no Experimento 3, para cinco termos como proposto nesta rodada de experimentos. Utilizando cinco termos o experimento obteve uma medida MAP de 11,01% diminuindo ainda mais o desempenho já pouco significativo alcançado no Experimento 3 que, para a mesma medida alcançou 16,74%. Uma explicação

para tal comportamento, apresentada também no Experimento 3 , pode ser o fato de que termos são muito genéricos e estão presentes em muitos documentos irrelevantes, e a sua utilização no contexto desse trabalho para RI resultou na recuperação de um número maior de documentos irrelevantes para a consulta em comparação aos experimentos 1 e 2 respectivamente. Ao analisarmos consulta a consulta os termos utilizados na sua expansão podemos constatar que: (i) foram utilizados 705 termos para a expansão das consultas; (ii) do total 113 (15,06%) termos foram considerados relevantes; e (iii) 637 (84,94%) termos irrelevantes.

6.8.6 Resultados do Experimento 6 com RR

Na Tabela 9 e na Figura 21 é possível observar que o aumento do número de RLBs não resulta na melhora do desempenho da recuperação de informação quando comparado com os demais experimentos que utilizaram RLBs apresentados nesta seção. De fato ao se utilizar as 10 RLBs mais pesadas dos três documentos melhor classificados pela recuperação considerados relevantes à consulta original, este obteve um valor para a medida MAP de 80,87%, sendo este o valor idêntico ao alcançado pelo Experimento 1, quer seja utilizando PRR e RR, onde foram utilizadas as três RLBs para os mesmos três documentos melhor classificados para a consulta original. Aliado a isso, o fato de que das 1500 RLBs utilizadas, somente 268 (17,87%) foram consideradas relevantes e 1233 (82,13%) não relevantes. Esta avaliação nos mostra que o percentual de RLBs relevantes utilizadas para a expansão das consultas é muito parecido ao percentual de RLBs utilizadas no Experimento 1 (20,88% utilizando PRR e 20,22% com RR), mesmo que em um número muito maior. Assim confirmamos que o aumento do número de RLBs na expansão das consultas não determina um aumento no desempenho da recuperação de informação se estas RLBs não for em um número expressivo de RLBs relevantes para as consultas.

6.9 Experimento com a exclusão das RLBs oriundas do Modelo TR+

Descrição: O quarto experimento avalia o impacto das relações lexicais binárias na recuperação de documentos. Para tanto esse experimento não adiciona novas RLBs à consulta original, e sim extrai as relações lexicais binárias definidas pelo Modelo TR+ para cada consulta.

O experimento tem seu início com a definição das consultas realizadas por Gonzalez, o próximo passo é a exclusão das RLBs oriundas de cada consulta. Em seguida a consulta é novamente realizada desta vez sem as RLBs.

Este experimento não necessita do processo de normalização (apresentados na Seção 6.2) uma vez que não são acrescidos à consulta original termos e relações lexicais binárias.

Objetivo: Com isso buscamos determinar o quão importantes são as RLBs vindas da con-

sulta original para a recuperação dos documentos.

Na Tabela 11 e na Figura 23, podemos observar o resultado da precisão e da abrangência do experimento realizado para avaliar a imporância das RLBs para a recuperação de informação.

Tabela 11 – Resultado do experimento com a exclusão das RLBs da consulta original

abrangência	precisão
0	0,9650
0,1	0,9551
0,2	0,9253
0,3	0,8971
0,4	0,8549
0,5	0,8411
0,6	0,7732
0,7	0,7434
0,8	0,6840
0,9	0,5420
1	0,3924
MAP	0,7778

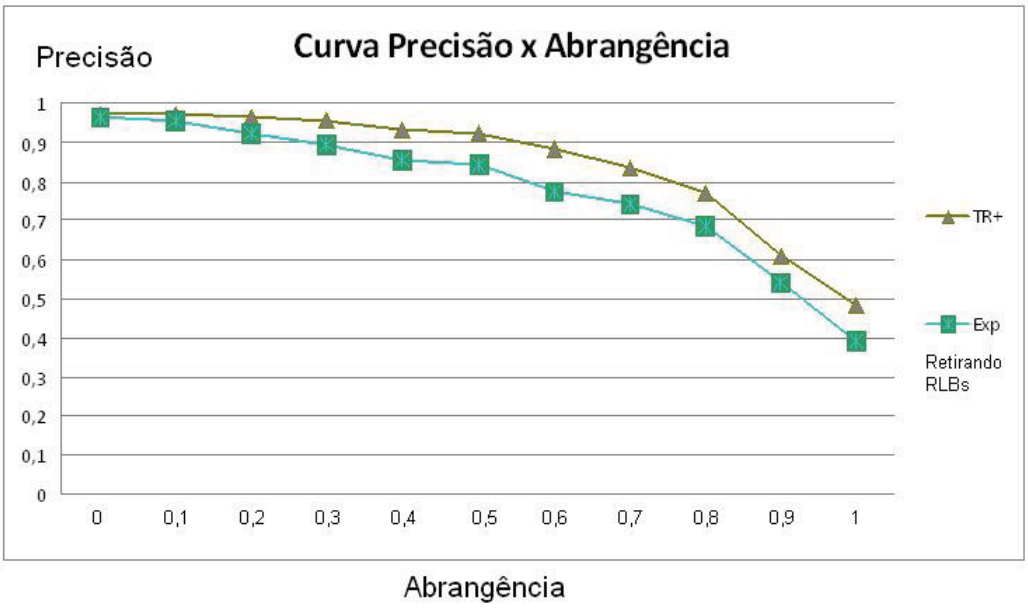

Figura 23 – Curva Precisão x Abrangência do experimento com a exclusão das RLBs da consulta original

Este experimento mostrou a importância das RLBs para a recuperação de informação, uma vez que ao se retirar as RLBs das consultas se obteve resultado inferior ao que foi atingido nos experimentos realizados por Gonzalez (2005). De fato ao se avaliar o desempenho do Modelo TR+ no que tange a recuperação de informação, este obteve para a medida MAP um percentual de 85,09% contra 77,78% quando são retiradas das consultas as RLBs. Ou seja, a EC não tradicional realizada pelo Modelo TR+ ao se utilizar de RLBs para a representação da

consulta original, esta constitui-se na expansão de consulta de melhor benefício no contexto desse trabalho.

6.10 Considerações sobre o capítulo

Neste capítulo apresentamos os experimentos realizados para a avaliação da proposta de EC realizada nesta dissertação. Apresentamos doze experimentos para as duas técnicas de EC utilizada nesse trabalho, Realimentação de Relevantes e Pseudo Realimentação de Relevantes, também apresentamos um experimento sem a aplicação das técnicas de expansão de consulta PRR e RR e sim a exclusão das RLBs utilizadas pela proposta do Modelo TR+. Apresentamos ainda para as duas técnicas (RR e PRR) o processo de normalização dos pesos, tanto das RLBs como dos termos envolvidos na EC junto ao Modelo TR+.

Os melhores resultados tanto para os experimentos com PRR quanto com RR foram alcançados, utilizando 3 RLBs e 10 RLBs dos 3 documentos melhor classificados (experimentos com PRR)ou dos 3 documentos escolhidos entre os 10 melhor classificados (experimentos com RR). Entretanto estes experimentos não superaram os resultados obtidos por Gonzalez (2005) que foram utilizados por nós como *baseline* para este trabalho.

Outra constatação sobre os experimentos é que, no contexto dessa dissertação, a utilização de termos tanto para PRR quanto para RR, mostrou-se ineficiente no que tange a RI.

Ao compararmos os experimentos tanto com PRR quanto com RR, podemos constatar que a técnica RR foi superior no que tange a medida MAP em comparação à PRR somente no Experimento 5. Esta semelhança ocorreu, devido ao fato de que, os documentos utilizados ((i) 3 documentos melhor ranqueados após a recuperação para PRR e (ii) 3 documentos melhor classificados pela recuperação e julgados como relevantes pelos usuários para RR) tanto para PRR como para RR foram muito semelhantes, ou seja praticamente os mesmos, diferenciando-se apenas em 8 consultas.

Com os resultados tão parecidos para os experimentos tanto com PRR como com RR, realizamos o Teste-T (Base, 2006) (veja Apendice D), para identificar a significância estatística dos resultados dos experimentos. Com o Teste-T podemos comparar os resultados dos experimentos com EC e o resultado do *baseline* (Gonzalez, 2005). Também utilizamos o Teste-T para compararmos os resultados dos experimentos com EC entre si, quer sejam com PRR e com RR.

Com o resultado do Teste-T podemos tecer algumas conclusões. Para os experimentos com PRR: (i) o *baseline* é superior estatisticamente ao Experimento 1, confirmando os resultados obtidos pela medida MAP para ambos; (ii) apesar do *baseline* ter atingido um valor para a medida MAP superior ao alcançado pelo Experimento 2 e suas variantes, o *baseline* não possui uma superioridade estatística significante em relação ao Experimento 2; (iii) o Teste-T aponta que o *baseline* é superior ao Experimento 4, o que confirma os valores obtidos por ambos pela medida MAP; (iv) apesar do *baseline* ter alcançado um valor superior para a medida MAP em

comparação ao Experimento 6, o Teste-T indica que não há uma superioridade estatística entre os dois experimentos; (v) ao compararmos os experimentos realizados com PRR que utilizaram RLBs para a EC entre si podemos concluir que, o Experimento 1 é superior estatisticamente aos Experimentos 4 e 6, e o Experimento 4 é superior estatisticamente ao Experimento 6, embora os três possuam o mesmo valor para a medida MAP; (vi) ao utilizarmos o Teste-T para comparar os resultados das variantes do Experimento 2 entre si, podemos concluir que não há diferença estatística significante entre as variações do Experimento 2; (vii) ao compararmos os resultados dos Experimentos 3 e 5 utilizando o Teste-T podemos concluir que não existe diferença significativa entre os dois experimentos. Para os experimentos com RR: (i) o *baseline* é superior estatisticamente ao Experimento 1 confirmando os valores da medida MAP de ambos; (ii) embora o *baseline* tenha alcançado um valor para a medida MAP superior às variantes do Experimento 2, ao compararmos estes resultados utilizando o Teste-T, o *baseline* não possui uma superioridade significativa as variações do Experimento 2; (iii) o Teste-T para os resultados do *baseline* e do Experimento 4 apontam pela superioridade estatística do *baseline*, confirmando os valores obtidos por ambos para a media MAP; (iv) apesar do *baseline* ter alcançado para a medida MAP um valor maior que o Experimento 6, ao compararmos estes dois experimentos utilizando o Teste-T, este indica que não há diferença estatística significativa entre eles; (v) ao compararmos os resultados das variantes do Experimento 2 utilizando o Teste-T podemos concluir que não há diferença estatística entres os três experimentos, como aponta os valores deles para a medida MAP; (vi) ao compararmos os resultados dos experimentos 1, 4 e 6 entre si (experimentos que utilizam RLBs para a EC) utilizando o Teste-T podemos concluir que, o Experimento 1 é superior ao 4 e 6, o Experimento 4 é superior ao Experimento 6, isto ocorre mesmo que para a medida MAP dos três experimentos tenham obtidos o mesmo valor; (vii) ao compararmos os experimentos 3 e 5 (experimentos que utilizam termos para a EC) utilizando o Teste-T podemos concluir que não há diferença estatística significante entre estes experimentos, ao compararmos entre si.

No próximo capítulo (Capítulo 7) apresentamos as conclusões sobre o trabalho desenvolvido nesta dissertação. Na Seção 7.2 apresentamos os resultados obtidos, e a públicação resultante do trabalho realizado. Na Seção 7.3 expomos as limitações encontradas no decorrer da dissertação e finalizando na Seção 7.4 apresentamos algumas sujestões para a continuidade do trabalho neste contexto.

7 Conclusões

7.1 Contextualização

Sistemas de Recuperação de Informação, que trabalham com documentos textuais, possuem como principal objetivo atender a consultas realizadas por usuários através de indexação, busca e classificação de documentos (Baeza-Yates & Ribeiro-Netto, 1999).

A maior dificuldade enfrentada pelo usuário, quanto à formulação adequada da consulta, é a decisão de quais palavras-chave usar para encontrar os documentos que necessita. Uma formulação eficiente passa pelo conhecimento do usuário sobre o domínio do tema a ser recuperado e sobre o próprio funcionamento do sistema. Entretanto, formular uma consulta eficiente através de palavras-chave, que possibilitem retornar informações relevantes, pode não ser uma tarefa fácil. Segundo Baeza-Yates e Ribeiro-Netto (1999), a identificação da real necessidade do usuário é um processo muito complexo e pode ser a diferença entre uma recuperação eficiente e uma recuperação que não atende as suas necessidades. Uma alternativa é a utilização de Expansão de Consulta, que reformula a consulta original para melhorar seu desempenho.

Para a representação dos conceitos e dos termos presentes nos documentos, diversas alternativas têm sido desenvolvidas e algumas incluem técnicas de Processamento da Língua Natural (PLN). Neste sentido, Gonzalez (2005) apresentou um modelo para recuperação de informação denominado TR+. O Modelo TR+ alia métodos estatísticos a conhecimento lingüístico para indexar e recuperar textos em língua portuguesa. Ele utiliza termos e relações lexicais binárias como descritores de conceitos.

Neste contexto, inseriu-se o objetivo desse trabalho, que foi de aplicar as técnicas de EC, Realimentação e Pseudo Realimentação de Relevantes, em um sistema que utiliza, para indexar e recuperar documentos textuais, o modelo de recuperação de informação TR+. Assim esteve inserida no contexto desse trabalho a discussão de experimentos usados para validar da proposta de EC em conjunto com o TR+.

7.2 Resultados Obtidos

No Capítulo 6 apresentamos os experimentos planejados e aplicados, referentes a cada uma das duas técnicas de EC, Realimentação e Pseudo Realimentação de Relevantes. Foram executadas ao todo 7 rodadas de experimentos para a posterior discussão de resultados. Utilizamos,

como *baseline* para os experimentos, os resultados obtidos por Gonzalez (2005).

Analizando os resultados da execução dos experimentos, podemos constatar que nenhum dos resultados obtidos, seja com RR seja com PRR, mensurados pela medida MAP, superaram os valores utilizados como *baseline*. Entrentanto, podemos destacar alguns pontos interessantes que resultaram dos experimentos: (i) RLBs são mais eficientes à EC do que termos no contexto dessa dissertação, tanto nos experimentos com RR como com PRR; (ii) o aumento do número de RLBs utilizadas na EC não aponta uma melhora no desempenho da RI; (iii) a utilização somente de termos mostrou-se bastante ineficiente no que tange a RI; (iv) ao compararmos os resultados dos experimentos tanto com PRR como com RR, podemos observar que os resultados foram bastante semelhantes. Isto ocorreu devido à pequena diferença entre os documentos que foram utilizados para a EC em ambas as técnicas (3 documentos melhor colocados pela consulta original e 3 documentos melhor ranqueados julgados relevantes pelos usuários). Somente o Experimento 5 com RR alcançou melhores resultados que o mesmo experimento com PRR. Isto atesta que a abordagem utilizada nesta dissertação que foi de utilizar RLBs e Termos melhor classificados dos documentos escolhidos para a EC não se mostrou o melhor método no contexto desse trabalho. O que nos leva crer que a análise prévia a EC das RLBs e dos Termos que serão utilizados seja uma abordagem mais eficiente.

O trabalho desenvolvido nesta dissertação resultou em uma publicação no *VI Workshop on Information and Human Language Technology*, com o seguinte título:

- Recuperação de Informação: Expansão de Consulta por Pseudo Realimentação de Relevantes no Modelo TR+ (Borges et al., 2008).

Neste trabalho apresentamos a especificação de experimentos para aplicação da técnica de expansão de consulta com pseudo realimentação de relevantes ao Modelo TR+ em recuperação de informação. Além dos experimentos realizados com PRR, também apresentamos seus resultados. O artigo foi apresentado na forma de poster, o que possibilitou além da divulgação do trabalho, a interação com pesquisadores da área ocasionando a troca de conhecimento e discussões sobre o tema.

7.3 Limitações

O Modelo TR+ foi inicialmente instanciado à língua portuguesa, não tendo ainda sido testado para outras línguas. Esta característica é um obstáculo a ser ultrapassado, uma vez que as possibilidades de realização de experimentos em outros idiomas estão vinculadas à instanciação do TR+. Isto acontece pois existem poucos corpora com as características necessárias para utilização pelo Modelo TR+. Este fator limitou a realização de outros experimentos para avaliar a aplicação tanto da técnica Pseudo Realimentação de Relevantes quanto da Realimentação de Relevantes em um corpus com um número mais expressivo de documentos que desse condições

de tecer uma análise mais consolidada.

7.4 Trabalhos Futuros

Para a continuidade do trabalho apresentado nesta dissertação, faz-se necessária a aplicação dos experimentos em um corpus com um volume de documentos significativamente maior que o utilizado em nossos experimentos. A aplicação em um corpus maior nos possibilitaria avaliar de forma significativa a aplicação de EC com RR e PRR em conjunto com o Modelo TR+. Entretanto, a aplicação da proposta em um corpus com maior volume demandaria as seguintes situações: (i) disponibilidade desse corpus significativamente maior ao utilizado; (ii) prototipação de ferramentas mais robustas que dessem suporte ao Modelo TR+ para a automação das etapas de pré-processamento e etiquetagem; (iii) instanciação do Modelo TR+ para outros idiomas, permitindo a indexação desses documentos e a posterior recuperação. Também futuramente, o julgamento da relevância das RLBs e Termos que serão utilizados na expansão das consultas é uma abordagem que merece maior atenção para a continuidade do trabalho apresentado nesta dissertação

Referências

Attar, R., & Fraenkel, A. S. (1977). Local feedback in full-text retrieval systems. New York, NY, USA, *J. ACM, vol. 24, July,* 397–417.

Avancini, H., Lavelli, A., Sebastiani, F., & Zanoli, R. (2006). Automatic expansion of domain-specific lexicons by term categorization. New York, NY, USA, *ACM Trans. Speech Lang. Process., vol. 3, May*, 1–30.

Baeza-Yates, R., & Ribeiro-Netto, B. (1999). *Modern information retrieval.* USA: Addison Wesley.

Barbetta, P. A., Reis, M. M., & Bornia, A. C. (2004). *Introduction to modern information retrieval. São Paulo, SP, Brasil. Ed. Altas.*

Base, R. M. K. (2006). The t-test. http://www.socialresearchmethods.net/kb/stat_t.php. Acessado em: 06/03/2009.

Borges, T. B., Gonzalez, M., & de Lima, V. L. S. (2008). Recuperação de informação: Expansão de consulta por pseudo realimentação de relevantes no modelo tr+. *VI Workshop Information and Human Language Technology, Vila Velha, ES, Brasil. SBC.* (381-384).

Broadbent, R. E., Saunders, G. S., & Ekstrom, J. J. (2006). An infrastructure for the evaluation and comparison of information retrieval systems. *SIGITE '06: Proceedings of the 7th conference on Information technology education* (123-127). New York, NY, USA: ACM Press.

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (25-32). New York, NY, USA: ACM Press.

Callan, J. P., Croft, W. B., & Broglio, J. (1995). Trec and tipster experiments with inquery. Tarrytown, NY, USA, *Inf. Proc. Manage, Pergamon Press, May, vol. 31,* 327–343.

Chen, A., & Gey, F. C. (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Inf. Retr. Hingham, ,MA, USA, Kluwer A. P., vol 7 January*, 149–182.

Chirita, P., F. C., & Nejdl, W. (2007). Personalized query expansion for the web. *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (7-14). New York, NY, USA: ACM Press.

Croft, W. B., Cook, R., & Wilder, D. (1995). Providing government information on the internet: Experiences with thomas. *In The Second International Conference on the Theory and Practice of Digital Libraries. New York, NY, USA. ACM Press.*

Crouch, C. J., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (77-88). New York, NY, USA: ACM

Custis, T., & Al-Kofahi, K. (2007). A new approach for evaluating query expansion: query-document term mismatch. *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (575-582). New York, NY, USA: ACM Press.

Dillon, M., Ulmschneider, J., & Desper, J. (1983). A prevalence formula for automatic relevance feedback in boolean systems. *Inf. P. Manage. vol.19, Elsevier Science Chapel Hill,NS,USA*, 27–36.

Ferreira, A. B. d. H. (1999). *Dicionário aurélio eletrônico - século xxi*. Rio de Janeiro, RJ, Brasil: Nova Fronteira: Lexikon Informática.

Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, vol 35, 243–255, June, Oxford Univ. Press. Oxford, UK.

Gonzalez, M. (2005). *Termos e relacionamentos em evidência na recuperação de informações.* Doctoral thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

Gonzalez, M., & de Lima, V. L. S. (2001). Semantic thesaurus for automatic expanded query in information retrieval. SPIRE (68-75), vol. 0. Lagura de San Rafael, Chile, November, IEEE Computer Science.

Gonzalez, M., de Lima, V. L. S., & de Lima, J. V. (2005). Binary lexical relations for text representation in information retrieval. NLDB (21-31). Springer, vol 10, n. 3, May, Berlim.

Gonzalez, M., de Lima, V. L. S., & de Lima, J. V. (2006a). Lexical normalization and relationship alternatives for a term dependence model in information retrieval.CICLing (394-405) Springer, Vol. 3878, n.7. January, Berlim.

Gonzalez, M. A. I., de Lima, V. L. S., & de Lima, J. V. (2006b). Tools for nominalization: An alternative for lexical normalization. PROPOR (100-109), vol 0. May, Springer. Berlim.

Harman, D. (1992). Relevance feedback revisited. *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (1-10). New York, NY, USA: ACM Press.

Houaiss, A. (2002). *Dicionário eletrônico houaiss da língua portuguesa: Versão 1.0.5.* Rio de Janeiro,Brasil: Objetiva.

Huang, X., Wen, M., An, A., & Huang, Y.-R. (2006). A platform for okapi-based contextual information retrieval. *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (728-738). New York, NY, USA: ACM Press.

Jing, Y., & Croft, W. B. (1994). *An association thesaurus for information retrieval* (Technical Report). University of Massachusetts. Amherst, MA, USA.

Jones, K. S. (1997). Search term relevance weighting given little relevance information. San Francisco, CA, USA, Morgam Kaufmann Publishers, 329–338.

Järvelin, K., & Kekäläinen, J. (2000). Ir, evaluation methods for retrieving highly relevant documents. *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (41-48). New York, NY, USA: ACM Press.

Kehdi, V. (2000). *Formação de palavras em português*. São Paulo,Br.: Atica.

Kowalski, G. (2000). *Information retrieval systems, theory and implementation, second edition.* Massachussetts,USA: Kluwer Academic Publisher.

Lee, K. S., Croft, W. B., & Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (235-242). New York, NY, USA: ACM Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press, Cambridged, UK.

Monz, C. (2003). *From document retrieval to question answering*. Doctoral thesis, University of Amsterdam Press, Amsterdam, Nederland.

Orengo, V. M. (2004). *Assessing relevance using automatically translated documents for cross-language information retrieval*. Doctoral thesis, Middlesex University. London, UK.

Orengo, V. M., & Huyck, C. (2006). Relevance feedback and cross-language information retrieval. Tarrytown, NY, USA, *Inf. Manage*, *vol. 42*, September 1203–1217. Pergamon Press.

Orengo, V. M., & Huyuck, C. (2002). Portuguese-english experiments using latent semantic *indexing. CLEF* (147-154), vol. 2785, February, Springer, Berlim.

Qiu, Y., & Frei, H.-P. (1993). Concept based query expansion. *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (160-169). New York, NY, USA: ACM Press.

Robertson, S. E., & Spark Jones, K. (1976). Relevance weighting of search terms. New York, NY, USA, *Journal of the American Soc. for Inf. Sciences*,*vol. 27, May*, 129–146,ACM Press.

S. E. Robertson, C. J. v. R., & Porter, M. F. (1981). Probabilistic models of indexing and searching. *SIGIR '80: Proceedings of the 3rd annual ACM conference on Research and development in information* (35-56). Kent, UK: Butterworth & Co.

Salton, G. (1971). *The smart retrieval system - experiments in automatic document processing.* Englewood, NJ, USA: Prentice Hall Inc.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *vol.24,March,Cornerl Univ., Ithaca,USA*, 513–523.

Salton, G., & Buckley, C. (1997). Improving retrieval performance by relevance feedback. San Francisco, CA, USA, 355–364, Morgan Kaufmann Publishers.

Salton, G., & MacGill, M. (1983). *Introduction to modern information retrieval.* New York, NY, USA: McGraw - Hill.

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the web: the public and their queries. New York, NY, USA, *J. A. Soc. Inf. Sci. Technol.*,*vol.52*,February,226–234.

Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (11-18). New York, NY, USA: ACM Press.

Vechtomova, O., & Karamuftuoglu, M. (2007). Query expansion with terms selected using lexical cohesion analysis of documents. Tarrytown, NY, USA, *Inf. Proc. Manage. vol 43*, July, 849–865, Pergamon Press.

Voorhees, E. M. (2005). The trec robust retrieval track. New York, NY, USA, *SIGIR Forum*, *39*, 11–20, ACM Publishers.

White, R. W., & Marchionini, G. (2006). A study of real-time query expansion effectiveness. *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (715-716). New York, NY, USA: ACM Press.

Wordnet (2008). Wordnet, lexical database of english. http://WordNet.princeton.edu/. Acessado em: 03/05/2007.

Xu, J., & Croft, B. W. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval(SIGIR)*. ACM. New York, NY, USA, (4-11).

Zhou, Y., & Croft, W. B. (2005). Document quality models for web ad hoc retrieval. *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management* (331-332). New York, NY, USA: ACM Press.

APÊNDICE A - Documentos utilizados para a EC com Pseudo Realimentação de Relevantes

- **Tópico 1:** *Abuso sexual*
  Documentos:407; 478; 437
- **Tópico 2:** *Acidente rodoviário*
  Documentos: 307; 308; 260
- **Tópico 3:** *Almoço*
  Documentos: 3014; 3028; 356
- **Tópico 4:** *Animação*
  Documentos: 2173; 2172; 1950
- **Tópico 5:** *Bolsa de valores*
  Documentos: 76; 643; 2790
- **Tópico 6:** *Campanha eleitoral de Lula*
  Documentos: 2997; 3505; 933
- **Tópico 7:** *Caso de cólera*
  Documentos: 354; 1317; 444
- **Tópico 8:** *Certificação*
  Documentos: 3571; 701; 1485
- **Tópico 9:** *Cinema brasileiro*
  Documentos: 1675; 1704; 3271
- **Tópico 10:** *Cirurgia*
  Documentos: 1324; 809; 2842
- **Tópico 11:** *Dança*
  Documentos: 3858; 3437; 3835
- **Tópico 12:** *Deputado federal*
  Documentos: 896; 930; 916
- **Tópico 13:** *Desemprego*
  Documentos: 2693; 2710; 855
- **Tópico 14:** *Digitalização*
  Documentos: 2125; 2540; 2122
- **Tópico 15:** *Distribuição de renda*
  Documentos: 2320; 1871; 2654

- **Tópico 16:** *Drible*
  Documentos: 1375; 462; 1162

- **Tópico 17:** *Escola de samba*
  Documentos: 4011; 2676; 4023

- **Tópico 18:** *Exportação*
  Documentos: 588; 618; 589

- **Tópico 19:** *Financiamento agrícola*
  Documentos: 14, 88; 968

- **Tópico 20:** *Franquia*
  Documentos: 688; 535; 3534

- **Tópico 21:** *Globalização*
  Documentos: 727; 2687; 2584

- **Tópico 22:** *Guerra do Golfo*
  Documentos: 844; 2241; 2892

- **Tópico 23:** *Hotel*
  Documentos: 3695; 1167; 3694

- **Tópico 24:** *Imóvel usado*
  Documentos: 1801; 1798; 1810

- **Tópico 25:** *Impressora*
  Documentos: 2018; 1986; 3555

- **Tópico 26:** *Informatização*
  Documentos: 1794; 1765; 1795

- **Tópico 27:** *Instrumento musical*
  Documentos: 1529; 3830; 2226

- **Tópico 28:** *Kit multimídia*
  Documentos: 2050; 2139; 2095

- **Tópico 29:** *Leilão de gado*
  Documentos: 8; 24; 106

- **Tópico 30:** *Liderança de campeonato*
  Documentos: 1252; 1009; 1195

- **Tópico 31:** *Medalha de ouro*
  Documentos: 3371; 864; 3420

- **Tópico 32:** *Merenda escolar*
  Documentos: 542; 273; 1642

- **Tópico 33:** *Mutuário*
  Documentos: 1822; 1805; 646

- **Tópico 34:** *Nudismo*
  Documentos: 3666; 2801; 3624

- **Tópico 35:** *Passeio de barco*
  Documentos: 3782; 3841; 3735

- **Tópico 36:** *Pastilha de freio*
  Documentos: 3293; 818; 4099

- **Tópico 37:** *Pintura restaurada*
  Documentos: 2264; 2787; 1634

- **Tópico 38:** *Plano real*
  Documentos: 234; 2379; 209

- **Tópico 39:** *Pólo turístico*
  Documentos: 3732; 3822; 3661

- **Tópico 40:** *Produtividade industrial*
  Documentos: 802; 701; 788

- **Tópico 41:** *Projeto arquitetônico*
  Documentos: 1900; 1771; 1832

- **Tópico 42:** *Propaganda eleitoral gratuita*
  Documentos: 931; 979; 915

- **Tópico 43:** *Publicação eletrônica*
  Documentos: 2074; 1498; 2071

- **Tópico 44:** *Reajuste salarial*
  Documentos: 646; 752; 770

- **Tópico 45:** *Reciclagem de lixo*
  Documentos: 571; 555; 572

- **Tópico 46:** *Seleção brasileira de futebol*
  Documentos: 1452; 1395; 1448

- **Tópico 47:** *Treino oficial*
  Documentos: 1206; 1252; 1036

- **Tópico 48:** *Uno Mille*
  Documentos: 4049; 4057; 4108

- **Tópico 49:** *Vestibular*
  Documentos: 1485; 2512; 1313

- **Tópico 50:** *Viagem de carro*
  Documentos: 4088; 3783; 4046

APÊNDICE B - Documentos utilizados para EC com Realimentação de Relevantes

- **Tópico 1:** *Abuso sexual*
  Documentos: 407; 478; 437
- **Tópico 2:** *Acidente rodoviário*
  Documentos: 307; 308; 260
- **Tópico 3:** *Almoço*
  Documentos: 3014; 3029; 356
- **Tópico 4:** *Animação*
  Documentos: 2173; 2172; 1950
- **Tópico 5:** *Bolsa de valores*
  Documentos: 2790; 651
- **Tópico 6:** *Campanha eleitoral de Lula*
  Documentos: 2927; 3505; 933
- **Tópico 7:** *Caso de cólera*
  Documentos: 354; 1317; 444
- **Tópico 8:** *Certificação*
  Documentos: 3571; 701; 1485
- **Tópico 9:** *Cinema brasileiro*
  Documentos: 1675; 1704; 3271
- **Tópico 10:** *Cirurgia*
  Documentos: 1324; 809; 2842
- **Tópico 11:** *Dança*
  Documentos: 3858; 3437; 3835
- **Tópico 12:** *Deputado federal*
  Documentos: 896; 930; 916
- **Tópico 13:** *Desemprego*
  Documentos: 2693; 2710; 855
- **Tópico 14:** *Digitalização*
  Documentos: 2125; 2540; 2122
- **Tópico 15:** *Distribuição de renda*
  Documentos: 2320; 1871; 2654

- **Tópico 16:** *Drible*
  Documentos: 1375; 462; 1162

- **Tópico 17:** *Escola de samba*
  Documentos: 4011; 2676; 4023

- **Tópico 18:** *Exportação*
  Documentos: 588; 618; 589

- **Tópico 19:** *Financiamento agrícola*
  Documentos: 14, 88; 87

- **Tópico 20:** *Franquia*
  Documentos: 688; 535; 3554

- **Tópico 21:** *Globalização*
  Documentos: 727; 2687; 2584

- **Tópico 22:** *Guerra do Golfo*
  Documentos: 844; 2241; 2892

- **Tópico 23:** *Hotel*
  Documentos: 3695; 1167; 3694

- **Tópico 24:** *Imóvel usado*
  Documentos: 1801; 1798; 1810

- **Tópico 25:** *Impressora*
  Documentos: 2018; 1986; 3555

- **Tópico 26:** *Informatização*
  Documentos: 1794; 1765; 1795

- **Tópico 27:** *Instrumento musical*
  Documentos: 1529; 3830; 2226

- **Tópico 28:** *Kit multimídia*
  Documentos: 2050; 2139; 2209

- **Tópico 29:** *Leilão de gado*
  Documentos: 8; 24; 106

- **Tópico 30:** *Liderança de campeonato*
  Documentos: 1252; 1009; 1195

- **Tópico 31:** *Medalha de ouro*
  Documentos: 3371; 864; 1382

- **Tópico 32:** *Merenda escolar*
  Documentos: 542; 273; 1642

- **Tópico 33:** *Mutuário*
  Documentos: 1822; 1805; 646

- **Tópico 34:** *Nudismo*
  Documentos: 3666; 2801; 3624

- **Tópico 35:** *Passeio de barco*
  Documentos: 3782; 3841; 3735

- **Tópico 36:** *Pastilha de freio*
  Documentos: 3293; 818; 4099

- **Tópico 37:** *Pintura restaurada*
  Documentos: 2264; 2787; 2778

- **Tópico 38:** *Plano real*
  Documentos: 234; 2379; 209

- **Tópico 39:** *Pólo turístico*
  Documentos: 3732; 3822; 3661

- **Tópico 40:** *Produtividade industrial*
  Documentos: 802; 701; 788

- **Tópico 41:** *Projeto arquitetônico*
  Documentos: 1900; 1771; 1832

- **Tópico 42:** *Propaganda eleitoral gratuita*
  Documentos: 931; 979; 915

- **Tópico 43:** *Publicação eletrônica*
  Documentos: 2074; 1498; 2071

- **Tópico 44:** *Reajuste salarial*
  Documentos: 646; 752; 770

- **Tópico 45:** *Reciclagem de lixo*
  Documentos: 571; 555; 572

- **Tópico 46:** *Seleção brasileira de futebol*
  Documentos: 1452; 1395; 1448

- **Tópico 47:** *Treino oficial*
  Documentos: 1206; 1252; 1036

- **Tópico 48:** *Uno Mille*
  Documentos: 4049; 4057; 4108

- **Tópico 49:** *Vestibular*
  Documentos: 1485; 2512; 1313

- **Tópico 50:** *Viagem de carro*
  Documentos: 4088; 3783; 4046

APÊNDICE C - Avaliação da relevância de documentos recuperados pelo Modelo TR+ com EC

Para a avaliação da relevância dos documentos recuperados foi elaborado um instrumento de avaliação da sua relevância de acordo com o **Tópico**, a **Descrição** e a **Narrativa**.

A seguir apresentamos um exemplo da apresentação de um documento avaliado por um usuário.

Nome:

O objetivo dessa avaliação é descobrir a relevância do documento julgado para cada consulta realizada. O avaliador deve julgar a relevância de cada documento levando em conta o **Tópico**, a **Descrição** e a **Narrativa** de cada tópico de consulta, marcando como relevante ou não relevante.

Tópico: Almoço
Descrição: Recuperar informação sobre almoço.
Narrativa: Um documento relevante deve relatar ou comentar encontros de pessoas para almoço ou informar sobre pratos servidos em um almoço ou, ainda, sobre preços ou locais deste tipo de refeição.

<461>Evite dar gafes em jantar japonês.<461>

Boas maneiras permitem pegar o sushi com a mão e recomendam coloca lo inteiro em a boca . Se você jamais for a o Japão , nem nunca se sentar a a mesa com um japonês , ainda=assim vale a pena guardar estas dicas : elas certamente farão mais feliz o seu sushiman preferido . A etiqueta japonesa não difere de a ocidental em seu princípio e função : ela serve para facilitar o trabalho de quem come e valorizar a refeição . Certas gafes frequentemente cometidas a a mesa oriental podem representar mais do=que um simples escorregão diplomático : elas podem interferir em a apreciação de o prato (veja quadro a o lado) . Por isso , no=caso=de os bolinhos de arroz , a regra número um é molha los sempre por a parte de o peixe , e não por a base . De a mesma maneira , deve se comer o sushi colocando o inteiro em a boca - usando a mão ou os palitos , tanto faz . Para comer sashimi , o correto é colocar pouco shoyu em o prato e não misturar em ele o wasabi (veja o=que significam os termos a o lado) . Esse tempero deve ser colocado sobre o peixe , para que , a o morder , a pessoa sinta todos=os sabores inteiros : o de a carne , o de o molho e o de o wasabi , explica Lumi Toyoda , professora de etiqueta japonesa . O sashimi é recomendável como entrada. Toshihiko Kumakura , diretor-presidente de o grupo Suntory em o Brasil , diz que a sucessão de pratos só é rigorosa em o banquete japonês tradicional , o kaiseki . Em essa ocasião, serve se primeiro um prato frio , com elementos de a estação e depois o sashimi . A etiqueta japonesa é de as mais fáceis de seguir - não há pecado que um=pouco de bom senso por=si não evite. Por=exemplo : jamais peça catchup a o sushiman . .

Em relação ao tópico, este documento é relevante: SIM( ) NÃO( )

APÊNDICE D - Análise estatística dos resultados dos experimentos realizados

Para uma análise mais aprofundada dos resultados obtidos nesta dissertação realizamos o cálculo denominado "Teste-T" (Base, 2006). O Teste-T é utilizado para compararmos as médias de duas amostragens e verificar se estas são estatísticamente diferentes entre elas. Esta análise torna-se importante uma vez que os resultados obtidos nos experimentos realizados foram muito parecidos entre si. No contexto dessa avaliação, analisamos a média da precisão para as 50 consultas[1] realizadas em todas as rodadas dos experimentos com PRR, RR e também para o *baseline* (Gonzalez, 2005). O cálculo para o "Teste-T" é realizado seguindo a fórmula (Barbetta et al., 2004):

$$t = \frac{\bar{d}.\sqrt{n}}{s_d} \tag{D.1}$$

onde:

- n é o tamanho da amostra[2];
- $\bar{d}$ é a média das diferenças observadas[3]; e
- s_d é o desvio padrão das diferenças observadas.

Neste apêndice apresentamos a precisão de cada uma das consulta realizadas para os experimentos tanto com PRR, como com RR e do *baseline* (Gonzalez, 2005), e também a média de todas as precisões para as 50 consultas.

D.1 Análise estatística utilizando o Teste-T para os experimentos com PRR

Na Tabela 12, apresentamos a precisão para todas as consultas dos experimentos com PRR que utilizaram de RLBs e termos para a expansão das consultas, também apresentamos a precisão para cada consulta do *baseline*.

Na Tabela 13 apresentamos os resultados para o Teste-T envolvendo os experimentos realizado por Gonzalez (2005) (*baseline*) e o Experimento 1 que utilizou as 3 RLBs melhor classificadas para os três documentos melhor classificados pela recuperação inicial.

Podemos observar na Tabela 13 que o valor para P(T<=t) bi-caudal (0,0162) é menor que o nível de significância padrão utilizado que é de 0,05, isto indica que o resultado do experimento realizados com o TR+ (*baseline*) é significativamente melhor que o resultado alcançado pelo Experimento 1.

[1] A média das precisões das 50 consultas é chamada neste trabalho de "*AvgPr*"

[2] Nos experimentos o tamanho da amostra consiste em 50 consultas.

[3] Nos experimentos a média das diferenças observadas consiste na média da precisão das consultas

Tabela 12 – Precisão para cada uma das 50 consultas dos experimentos utilizando RLBs e Termos com PRR

	TR+	Exp 1	Exp 2.1	Exp 2.2	Exp 2.3	Exp 3	Exp 4	Exp 5	Exp 6
Consulta	Pr	Pr	Pr	Pr	Pr	Pr	Pr	Pr	Pr
301	0,8065	0,8065	0,8065	0,8065	0,8065	0,323	0,8065	0,0000	0,8065
302	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
303	0,9792	0,9038	0,9792	0,9792	0,9792	0,1250	0,9792	0,1667	0,9792
304	0,6457	0,3457	0,3457	0,3457	0,3457	0,1235	0,3457	0,1358	0,3457
305	0,1250	0,1250	0,1250	0,1250	0,1250	0,0000	0,1250	0,0000	0,1250
306	0,7368	0,6829	0,7368	0,8684	0,8684	0,1316	0,7368	0,1053	0,9091
307	0,9091	0,9091	0,9091	0,9091	0,9091	0,8112	0,9091	0,0000	0,9091
308	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
309	0,5000	0,5000	0,5000	0,5000	0,5000	0,0000	0,5000	0,0000	0,5000
310	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
311	0,8824	0,8824	0,8824	0,8824	0,8824	0,5882	0,8824	0,5735	0,8824
312	0,8462	0,7674	0,8462	0,8462	0,8462	0,1026	0,8462	0,1538	0,8462
313	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0741	1,0000
314	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
315	0,6000	0,3750	1,0000	0,6000	0,6000	0,0000	0,6000	0,0000	0,6000
316	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
317	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
318	0,8873	0,8873	0,8873	0,8873	0,8873	0,8592	0,8873	0,4507	0,8873
319	0,7273	0,7500	0,7273	0,7273	0,7273	0,0000	0,7273	0,0909	0,7273
320	0,9286	0,9286	0,9286	0,9286	0,9286	0,0357	0,9286	0,0357	0,9286
321	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
322	1,0000	0,9167	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
323	0,8600	0,8600	0,8600	0,8600	0,8600	0,8300	0,8600	0,8300	0,8600
324	0,8333	0,8571	0,8333	0,8333	0,8333	0,0000	0,8333	0,0000	0,8333
325	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
326	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
327	0,7959	0,7959	0,7959	0,7959	0,7959	0,0000	0,7959	0,0000	0,7959
328	0,1111	0,1111	0,1111	0,1111	0,1111	0,0000	0,7500	0,0000	0,1111
329	0,7500	0,7500	0,7500	0,7500	0,7500	0,0000	0,7500	0,0000	0,7500
330	0,6786	0,5135	0,6786	0,6786	0,6786	0.0714	0,7500	0,0000	0,6786
331	0,7500	0,7500	0,7500	0,7500	0,7500	0,0000	1,0000	0,1429	0,7500
332	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,8000	0,0000	1,0000
333	0,8000	0,8000	0,8000	0,8000	0,8000	0,6000	1,0000	0,0000	0,8000
334	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,8000	0,0000	1,0000
335	0,8000	0,8000	0,8000	0,8000	0,8000	0,0000	1,0000	0,0000	0,8000
336	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,6667	0,0000	1,0000
337	0,6667	0,6667	0,6667	0,6667	0,6667	0,0000	0,9464	0,0000	0,6667
338	0,9464	0,9322	0,9464	0,9464	0,9464	0,3393	0,8000	0,0000	0,9464
339	0,8000	0,2857	0,8000	0,8000	0,8000	0,0000	0,8462	0,3036	0,8000
340	0,8462	0,5909	0,8462	0,8462	0,8462	0,0000	0,7143	0,0000	0,8462
341	0,7143	0,6250	0,7143	0,7143	0,7143	0,0000	0,0000	0,0000	0,7143
342	0,5714	0,5714	0,5714	0,5714	0,5714	0,7143	0,0000	0,0000	0,5714
343	0,8750	0,8750	0,8750	0,8750	0,8750	0,0000	0,0000	0,0000	0,8750
344	0,5385	0,5385	0,5385	0,5385	0,5385	0,1154	0,0000	0,1154	0,5385
345	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,0000	0,0000	1,0000
346	0,6575	0,6575	0,6575	0,6575	0,6575	0,6712	0,0000	0,5890	0,6575
347	0,9167	0,8462	0,9167	0,9167	0,9167	0,0833	0,0000	0,0833	0,9167
348	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,0000	0,0000	1,0000
349	0,9583	0,9583	0,9583	0,9583	0,9583	0,0000	0,0000	0,0000	0,9583
350	0,8174	0,8182	0,8182	0,8182	0,8182	0,0000	0,0000	0,0000	0,8182
AvgPr	0,8174	0,7859	0,8254	0,8201	0,8201	0,1248	0,6685	0,0770	0,8174

Tabela 13 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 1 com PRR

	TR+	Exp 1
Média	0,8174	0,7858
Variância	0,0458	0,0557
Observações	50	50
Correlação de Pearson	0,9253	
Hipótese da diferença média	0	
gl	49	
Stat t	2,4891	
P(T<=t) uni-caudal	0,0081	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0162	
t crítico bi-caudal	2,0095	

Tabela 14 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 2.1 com PRR

	TR+	Exp 2.1
Média	0,817	0,8200
Variância	0,0458	0,0457
Observações	50	50
Correlação de Pearson	0,9962	
Hipótese da diferença média	0	
gl	49	
Stat t	-1,0061	
P(T<=t) uni-caudal	0,1596	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,3192	
t crítico bi-caudal	2,0095	

Nas tabelas 14, 15 e 16 apresentamos os resultados para o Teste-T comparando os resultados dos experimentos utilizados como *baseline* (Gonzalez, 2005) e as variações do Experimento 2 (utilizando as 3 RLBs melhor classificadas para os três documentos melhor classificados pela recuperação inicial, sendo utilizadas os três tipos de RLBs: (i) Restrição; (ii) Classificação e (iii) Associação).

Nas tabelas 14, 15 e 16 podemos observar, que o valor obtido para P(T<=t) bi-caudal (Tabela 14 = 0,3192; Tabela 15 = 0,3196 e Tabela 16 = 0,3192) é maior que o valor padrão de significância (0,05). Isto nos leva a deduzir, que apesar do experimento realizado com o TR+ (*baseline*) ter alcançado valor para a medida MAP superior aos obtidos pelos experimento 2.1; 2.2 e 2.3 para a mesma medida, não há diferença significante entre o *baseline* e o Experimento 2 e suas variantes quando realizado o Teste-t.

Na Tabela 17 apresentamos os resultados para o Teste-T envolvendo a comparação dos resultados obtidos pelo *baseline* e os resultados alcançados no Experimento 4, utilizando 5 RLBs melhor classificadas dos três documentos melhores classiificados pela recuperação inicial.

Podemos observar na Tabela 17 que o valor para P(T<=t) bi-caudal (0,0057) é inferior ao nível de significância padrão (0,05) utilizado para este teste, isto significa que os resultados obtidos pelo TR+ é siginificativamente superior aos resultados alcançados pelo Experimento 4.

Na Tabela 18 apresentamos os resultados para o Teste-T envolvendo a comparação dos resultados obtidos pelo *baseline* e os resultados alcançados no Experimento 6, utilizando 10 RLBs melhor classificadas dos três documentos melhores classiificados pela recuperação inicial.

Na Tabela 18 é possivel obervar que o valor para P(T<=t) bi-caudal (0,3222) é maior que o valor de significância padrão (0,05) utilizado neste teste. Isto significa que o TR+ com seus

Tabela 15 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 2.2 com PRR

	TR+	Exp 2.2
Média	0,8174	0,8200
Variância	0,0458	0,0457
Observações	50	50
Correlação de Pearson	0,9962	
Hipótese da diferença média	0	
gl	49	
Stat t	-1,0061	
P(T<=t) uni-caudal	0,1596	
t crítico uni-caudal	1,676	
P(T<=t) bi-caudal	0,3196	
t crítico bi-caudal	2,0095	

Tabela 16 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 2.3 com PRR

	TR+	Exp 2.3
Média	0,8174	0,8200
Variância	0,0458	0,0457
Observações	50	50
Correlação de Pearson	0,9962	
Hipótese da diferença média	0	
gl	49	
Stat t	-1,0061	
P(T<=t) uni-caudal	0,1596	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,3192	
t crítico bi-caudal	2,0095	

Tabela 17 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 4 com PRR

	TR+	Exp 4
Média	0,8178	0,6685
Variância	0,0458	0,1439
Observações	50	50
Correlação de Pearson	0,9962	
Hipótese da diferença média	0	
gl	49	
Stat t	-1,0061	
P(T<=t) uni-caudal	0,0028	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0057	
t crítico bi-caudal	2,0095	

Tabela 18 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 6 com PRR

	TR+	Exp 6
Média	0,8174	0,8174
Variância	0,0458	0,0458
Observações	50	50
Correlação de Pearson	0,9999	
Hipótese da diferença média	0	
gl	49	
Stat t	-1	
P(T<=t) uni-caudal	0,1611	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,3222	
t crítico bi-caudal	2,0095	

Tabela 19 – Teste-t: duas amostras em par para médias para os experimentos Exp 1 e Exp 4 com PRR

	Exp 1	Exp 4
Média	0,7858	0,6685
Variância	0,0557	0,1439
Observações	50	50
Correlação de Pearson	0,3036	
Hipótese da diferença de média	0	
gl	49	
Stat t	2,1768	
P(T<=t) uni-caudal	0,0171	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0343	
t crítico bi-caudal	2,0095	

resultados não é significativamente melhor que o Experimento 6, apesar de que o Modelo TR+ tenha alcançado para a medida MAP 85,59& contra 80,87% do Experimento 6 para a mesma medida.

Nas tabelas 19, 20 e 21 apresentamos os resultados para o Teste-T, comparando os resultados obtidos pelos experimentos 1, 4 e 6 entre si. Esta análise foi realizada pelo fato que os três experimentos citados utilizam 3, 5 e 10 RLBs dos 3 documentos melhor classificados pela recuperação inicial, independentemente do seu tipo.

Podemos observar nas tabelas 19, 20 e 21, os resultados para o Teste-t entre os experimentos 1, 4 e 6 entre si. Na Tabela 19 temos que o valor para P(T<=t) bi-caudal (0,0161) é menor que o valor de significância (0,05) utilizado neste teste. Isto significa que o Experimento 1 é significativamente melhor no que tange aos seus resultados do que o Experimento 4, confirmando assim os valores para a medida MAP que foi de 80,87% e 65,26% respectivamente. Na Tabela 20 podemos observar que o valor para P(T<=) bi-caudal (0,0161) é inferior ao valor de significância (0,05) utilizado neste teste. Isto indica que, mesmo que ambos os experimentos tenham alcançados o valor para a medida MAP de 80,87%, o Experimento 1 é significativamente melhor que o Experimento 6. Na Tabela 21 podemos observar que o valor para P(T<=t) bi-caudal (0,0057) é inferior ao valor de significancia (0,05) utilizado neste teste. Podemos afirmar com isso que o mesmo o Experimento 4 é significativamente superior ao Experimento 6.

Nas tabelas 22, 23 e 24 apresentamos os resultados para o Teste-T, comparando os resultados das variações do Experimento 2 entre si.

Nas tabelas 24, 22 e 23 podemos observar que o valor para P(T<=t) bi-caudal (0,5293) nas 3 tabelas foram bem maiores do que o valor de significância (0,05) utilizado neste teste. Com

Tabela 20 – Teste-t: duas amostras em par para médias para os experimentos Exp 1 e Exp 6 com PRR

	Exp 1	Exp 6
Média	0,7858	0,6685
Variância	0,0557	0,0458
Observações	50	50
Correlação de Pearson	0,9253	
Hipótese da diferença de média	0	
gl	49	
Stat t	-2,4906	
P(T<=t) uni-caudal	0,0080	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0161	
t crítico bi-caudal	2,0095	

Tabela 21 – Teste-t: duas amostras em par para médias para os experimentos Exp 4 e Exp 6 com PRR

	Exp 1	Exp 6
Média	0,6685	0,0817
Variância	0,1439	0,0458
Observações	50	50
Correlação de Pearson	0,3500	
Hipótese da diferença de média	0	
gl	49	
Stat t	-2,8881	
P(T<=t) uni-caudal	0,0028	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0057	
t crítico bi-caudal	2,0095	

Tabela 22 – Teste-t: duas amostras em par para médias para os experimentos 2.1 e 2.2 com PRR

	Exp 2.1	Exp 2.2
Média	0,8254	0,8200
Variância	0,0454	0,0457
Observações	50	50
Correlação de Pearson	0,9606	
Hipótese da diferença de média	0	
gl	49	
Stat t	0,6335	
P(T<=t) uni-caudal	0,2646	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,5293	
t crítico bi-caudal	2,0095	

Tabela 23 – Teste-t: duas amostras em par para médias para os experimentos 2.2 e 2.3 com PRR

	Exp 2.2	Exp 2.3
Média	0,8254	0,8200
Variância	0,0454	0,0457
Observações	50	50
Correlação de Pearson	0,9606	
Hipótese da diferença de média	0	
gl	49	
Stat t	0,6335	
P(T<=t) uni-caudal	0,2646	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,5293	
t crítico bi-caudal	2,0095	

Tabela 24 – Teste-t: duas amostras em par para médias para os experimentos 2.1 e 2.3 com PRR

	Exp 2.1	Exp 2.2
Média	0,8254	0,8200
Variância	0,0454	0,0457
Observações	50	50
Correlação de Pearson	0,9606	
Hipótese da diferença de média	0	
gl	49	
Stat t	0,6335	
P(T<=t) uni-caudal	0,2646	
t crítico uni-caudal	1,676	
P(T<=t) bi-caudal	0,5293	
t crítico bi-caudal	2,0095	

Tabela 25 – Teste-t: duas amostras em par para médias para os experimentos Exp 3 e o Exp 5 com PRR

	Exp 3	Exp 5
Média	0,1248	0,0770
Variância	0,0650	0,0306
Observações	50	50
Correlação de Pearson	0,6999	
Hipótese da diferença de média	0	
gl	49	
Stat t	1,8560	
P(T<=t) uni-caudal	0,0347	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0694	
t crítico bi-caudal	2,0095	

isto podemos afirmar que não há diferença significativa entre as variações do experimento 2 utilizando PRR.

Na Tabela 25 apresentamos os resultados para o Teste-T comparando os resultados dos experimentos 3 e 5 entre si. Os experimentos 3 e 5 utilizam termos ao invéz de RLBs para a EC. Sendo que o Experimento 3 utiliza os 3 termos melhor classificados dos três primeiros documentos recuperados pela consulta inicial, e o Experimento 5 utiliza os 5 termos melhor classificados dos mesmos documentos.

Na Tabela 25 é possível observar que o valor a para P(T<=t) bi-caudal (0,0694) é superior ao valor de significância (0,05), o que nos possibilita concluir que não há diferença significativa entre os experimentos 3 e 5 utilizando PRR.

D.2 Análise estatística utilizando o Teste-T para os experimentos com RR

Nesta seção apresentamos a precisão de cada uma das consulta realizadas para os experimentos com RR e do *baseline* (Gonzalez, 2005), a média de todas as precisões para as 50 consultas (Tabela 26).

Na Tabela 27 apresentamos o resultado para o Teste-T, comparando os experimentos TR+ e 1 com RR. O Experimento 1 com RR utiliza as três RLBs mais com maior peso dos três primeiros documentos julgados relevantes pelo usuario de acordo com cada uma das 50 consultas.

Podemos observar na Tabela 27, que o resultado para P(T<=t) bi-caudal (0,0185) é inferior ao valor de significância padrão (0,005) utilizado neste teste. Iso nos indica que, o TR+ é significativamente superior ao Experimento 1 no que tange ao contexto deste trabalho.

Na Tabela 28 apresentamos o resultado para o Teste-T entre o Modelo TR+ e o Experimento 2.1. O Experimento 2.1 utiliza para a EC as 3 RLBs do tipo Restrição dos três primeiros documentos julgados como relevantes para a consulta original pelo usuário.

Na Tabela 28 podemos observar que o o valor para P(T<=t) bi-caudal (0,2960) é superior ao valor de significância padrão (0,05) utilizado neste teste. Com isto podemos concluir que não existe diferença significante entre o TR+ e p Experimento 2.1.

Na Tabela 29 apresentamos o resultado para o Teste-T entre o Modelo TR+ e o Experimento 2.2. O Experimento 2.2 utiliza para a EC as 3 RLBs do tipo Associação dos três primeiros documentos julgados como relevantes para a consulta original pelo usuário.

Na Tabela 29 podemos observar que o valor para P(T<=t) bi-caudal (0,2490) é superior ao valor de significância padrão (0,05) utilizado neste teste. Podemos concluir com isso que o TR+ não é significativamente superior ao Experimento 2.2 no que tange aos resultados dos

Tabela 26 – Precisão para cada uma das 50 consultas dos experimentos utilizando RLBs e Termos com RR

	TR+	Exp 1	Exp 2.1	Exp 2.2	Exp 2.3	Exp 3	Exp 4	Exp 5	Exp 6
Consulta	Pr	Pr	Pr	Pr	Pr	Pr	Pr	Pr	Pr
301	0,8065	0,8065	0,8065	0,8065	0,8065	0,0323	0,8065	0,0000	0,8065
302	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
303	0,9792	0,9184	0,9792	0,9792	0,9792	0,1375	0,9792	01695	0,9792
304	0,6457	0,3457	0,3457	0,3457	0,3457	0,1235	0,3457	0,1358	0,3457
305	0,1250	0,1375	0,1250	0,1250	0,1250	0,0000	0,1250	0,0000	0,1250
306	0,7368	0,6829	0,7368	0,8684	0,8684	0,1316	0,7368	0,1053	0,9091
307	0,9091	0,9091	0,9091	0,9091	0,9091	0,8182	0,9091	0,0000	0,9091
308	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
309	0,5000	0,5000	0,5000	0,5000	0,5000	0,0000	0,5000	0,0000	0,5000
310	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
311	0,8824	0,8824	0,8824	0,8824	0,8824	0,5882	0,8824	0,5735	0,8824
312	0,8462	0,7674	0,8462	0,8462	0,8462	0,1026	0,8462	0,1538	0,8462
313	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0741	1,0000
314	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
315	0,6000	0,3750	1,0000	0,6000	0,6000	0,0000	0,6000	0,0000	0,6000
316	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
317	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
318	0,8873	0,8873	0,8873	0,8873	0,8873	0,8592	0,8873	0,4507	0,8873
319	0,7273	0,7563	0,7273	0,7273	0,7273	0,0000	0,7273	0,0988	0,7273
320	0,9286	0,9286	0,9286	0,9286	0,9286	0,0357	0,9286	0,0357	0,9286
321	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
322	1,0000	0,9167	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
323	0,8600	0,8600	0,8600	0,8600	0,8600	0,8300	0,8600	0,8300	0,8600
324	0,8333	0,8571	0,8333	0,8333	0,8333	0,0000	0,8333	0,0000	0,8333
325	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
326	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	1,0000	0,0000	1,0000
327	0,7059	0,7059	0,7059	0,7059	0,7059	0,0000	0,7059	0,0000	0,7059
328	0,1101	0,1101	0,1101	0,1101	0,1101	0,0000	0,7500	0,0000	0,1101
329	0,7500	0,7500	0,7500	0,7500	0,7500	0,0000	0,7500	0,0000	0,7500
330	0,6786	0,5135	0,6786	0,6786	0,6786	0,0714	0,7500	0,1429	0,6786
331	0,7500	0,7500	0,7500	0,7500	0,7500	0,0000	1,0000	0,0000	0,7500
332	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,8000	0,0000	1,0000
333	0,8000	0,8000	0,8000	0,8000	0,8000	0,6000	1,0000	0,0000	0,8000
334	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,8000	0,0000	1,0000
335	0,8000	0,8000	0,8000	0,8000	0,8000	0,0000	1,0000	0,0000	0,8000
336	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,6667	0,0000	1,0000
337	0,6667	0,6637	0,6637	0,6637	0,6637	0,0000	0,9474	0,0000	0,6637
338	0,9464	0,9322	0,9464	0,9464	0,9464	0,3393	0,8000	0,3036	0,9464
339	0,8000	0,2857	0,8000	0,8000	0,8000	0,0000	0,8462	0,0000	0,8000
340	0,8462	0,5909	0,8462	0,8462	0,8462	0,0000	0,7143	0,0000	0,8462
341	0,7143	0,6250	0,7143	0,7143	0,7143	0,0000	0,0000	0,0000	0,7143
342	0,5714	0,5714	0,5714	0,5714	0,5714	0,7143	0,0000	0,0000	0,5714
343	0,8750	0,8750	0,8750	0,8750	0,8750	0,0000	0,0000	0,0000	0,8750
344	0,5385	0,5385	0,5385	0,5385	0,5385	0,1154	0,0000	0,1154	0,5385
345	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,0000	0,0000	1,0000
346	0,6575	0,6575	0,6575	0,6575	0,6575	0,6712	0,0000	0,5890	0,6575
347	0,9167	0,8462	0,9167	0,9167	0,9167	0,0833	0,0000	0,0833	0,9167
348	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000	0,0000	0,0000	1,0000
349	0,9583	0,9583	0,9583	0,9583	0,9583	0,0000	0,0000	0,0000	0,9583
350	0,8174	0,8182	0,8182	0,8182	0,8182	0,0000	0,0000	0,0000	0,8182
AvgPr	0,8174	0,7859	0,8254	0,8201	0,8201	0,1251	0,6685	0,0772	0,8174

Tabela 27 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 1 com RR

	TR+	Exp 1
Média	0,8172	0,7863
Variância	0,0458	0,0555
Observações	50	50
Correlação de Pearson	0,9248	
Hipótese da diferença de média	0	
gl	49	
Stat t	2,4347	
P(T<=t) uni-caudal	0,0092	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0185	
t crítico bi-caudal	2,0095	

Tabela 28 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 2.1 com RR

	TR+	Exp 2.1
Média	0,8172	0,8257
Variância	0,04587	0,0452
Observações	50	50
Correlação de Pearson	0,9649	
Hipótese da diferença de média	0	
gl	49	
Stat t	1,0562	
P(T<=t) uni-caudal	0,1480	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,2960	
t crítico bi-caudal	2,0095	

Tabela 29 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 2.2 com RR

	TR+	Exp 2.2
Média	0,8172	0,8203
Variância	0,04587	0,04548
Observações	50	50
Correlação de Pearson	0,9961	
Hipótese da diferença de média	0	
gl	49	
Stat t	-1,1666	
P(T<=t) uni-caudal	0,1245	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,2490	
t crítico bi-caudal	2,0095	

Tabela 30 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 2.3 com RR

	TR+	Exp 2.3
Média	0,8172	0,8203
Variância	0,0458	0,0454
Observações	50	50
Correlação de Pearson	0,9961	
Hipótese da diferença de média	0	
gl	49	
Stat t	-1,1666	
P(T<=t) uni-caudal	0,1245	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,2490	
t crítico bi-caudal	2,0095	

Tabela 31 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 4 com RR

	TR+	Exp 4
Média	0,8172	0,6684
Variância	0,0458	0,1440
Observações	50	50
Correlação de Pearson	0,3519	
Hipótese da diferença de média	0	
gl	49	
Stat t	2,8888	
P(T<=t) uni-caudal	0,0028	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0057	
t crítico bi-caudal	2,0095	

experimentos relizado neste trabalho.

Na Tabela 30 apresentamos o resultado para o Teste-T entre o Modelo TR+ e o Experimento 2.3. O Experimento 2.1 utiliza para a EC as 3 RLBs do tipo Classificação dos três primeiros documentos julgados como relevantes para a consulta original pelo usuário.

Na Tabela 30 podemos observar que o valor para P(T<=t) bi-caudal (0,2490) é superior ao valor padrão de significância (0,05) utilizado neste teste. Com isto podemos concluir que o TR+ não é significativamente melhor que o Experimento 2.3 no que tange ao contexto deste trabalho.

Na Tabela 31 apresentamos o resultado para o Teste-T entre o Modelo TR+ e o Experimento 4. O Experimento 4 utiliza para a EC as 5 RLBs dos três primeiros documentos julgados como relevantes para a consulta original pelo usuário.

Na Tabela 31 podemos observar que o valor para P(T<=t) bi-caudal (0,0057) é menor que o valor de significância padrão (0,05) adotado neste teste. Assim podemos concluir que o TR+ é significativamente superior ao Experimento 4 no que tange aos experimentos realizados neste trabalho.

Na Tabela 32 apresentamos o resultado para o Teste-T entre o Modelo TR+ e o Experimento 6. O Experimento 6 utiliza para a EC as 10 RLBs dos três primeiros documentos julgados como relevantes para a consulta original pelo usuário.

Na Tabela 32 observamos que P(T<=t) bi-caudal obteve um valor (0,1835) superior ao valor padrão de significância (0,05) adotado neste teste. Assim podemos afirmar que o TR+ não é significativamente melhor que o Experimento 6 quanto aos resultados obtidos neste trabalho.

Nas tabelas 33, 34 e 35 apresentamos o resultado para o Teste-T entre as variantes do Experimento 2 entre si.

Tabela 32 – Teste-t: duas amostras em par para médias para os experimentos TR+ e Exp 6 com RR

	TR+	Exp 4
Média	0,8172	0,8174
Variância	0,0458	0,0459
Observações	50	50
Correlação de Pearson	0,9999	
Hipótese da diferença de média	0	
gl	49	
Stat t	-1,3490	
P(T<=t) uni-caudal	0,0917	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,1835	
t crítico bi-caudal	2,0095	

Tabela 33 – Teste-t: duas amostras em par para médias para os experimentos Exp 2.1 e Exp 2.2 com RR

	Exp 2.1	Exp 2.2
Média	0,8257	0,8203
Variância	0,0452	0,0454
Observações	50	50
Correlação de Pearson	0,9604	
Hipótese da diferença de média	0	
gl	49	
Stat t	0,6335	
P(T<=t) uni-caudal	0,2646	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,5293	
t crítico bi-caudal	2,0095	

Tabela 34 – Teste-t: duas amostras em par para médias para os experimentos Exp 2.1 e Exp 2.3 com RR

	Exp 2.1	Exp 2.3
Média	0,8257	0,8203
Variância	0,0452	0,0454
Observações	50	50
Correlação de Pearson	0,9604	
Hipótese da diferença de média	0	
gl	49	
Stat t	0,6335	
P(T<=t) uni-caudal	0,2646	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,5293	
t crítico bi-caudal	2,0095	

Tabela 35 – Teste-t: duas amostras em par para médias para os experimentos Exp 2.2 e Exp 2.3 com RR

	Exp 2.2	Exp 2.3
Média	0,8172	0,1250
Variância	0,0458	0,0650
Observações	50	50
Correlação de Pearson	-,042667	
Hipótese da diferença de média	0	
gl	49	
Stat t	14,3986	
P(T<=t) uni-caudal	1,5569	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	3,1138	
t crítico bi-caudal	2,0095	

Tabela 36 – Teste-t: duas amostras em par para médias para os experimentos Exp 1 e Exp 4 com RR

	Exp 1	Exp 4
Média	07863	0,6684
Variância	0,0555	0,1444
Observações	50	50
Correlação de Pearson	0,3055	
Hipótese da diferença de média	0	
gl	49	
Stat t	2,1906	
P(T<=t) uni-caudal	0,0166	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0332	
t crítico bi-caudal	2,0095	

Podemos observar nas tabelas 35, 35 e 35 que os valores para P(T<=t) bi-caudal são maiores que o valor de significância padrão (0,05) que foi utilizado neste teste. Assim podemos concluir que as variações do Experimento 2 quando comparadas entre si, não apresentam diferenças significativas umas com as outras.

Na Tabela 36 apresentamos os resultados do Teste-T para a comparação dos resultados dos Experimentos 1 e 4. O Experimento 1 utiliza as 3 RLBs melhor classificadas dos três primeiros documentos julgados relevantes para a consulta original, enquanto o Experimento 4 utiliza, as RLBs com o mesmo critério de selecção do Experimento 1.

Podemos observar na Tabela 36 que o valor para P(T<=t) bi-caudal (0,0332) é inferior ao valor da significância padrão (0,05) utilizado neste teste, o que indica que o Experimento 1 é significativamente superior ao Experimento 4 no que tange este trabalho.

Na Tabela 37 apresentamos o resultado para o Teste-T ao compararmos os resultados obtidos pelos Experimentos 1 e 6. O Experimento 1 utiliza as três RLBs retiradas dos 3 primeiros documentos julgados relevantes para a consulta original, enquanto o Experimeto 6 utiliza as RLBs seleciondas pelo mesmo critério.

Podemos observar na Tabela 37 que o valor para P(T<=t) bi-caudal (0,0179) é inferior ao valor de significância padrão (0,05) utilizado neste teste. Com isto podemos concluir que o Experimento 1 é significativamente superior ao Experimento 6 no contexto deste trabalho.

Na Tabela 38 apresentamos o resultado do Teste-t envolvendo os experimentos 4 e 6. O Experimento 4 utiliza as 5 RLBs melhores classificadas dos três primeiros documentos julgados relevantes pelo usuário de acordo com a consulta original, já o Experimento 6 utiliza as 10 RLBs com o mesmo critério de seleção.

Tabela 37 – Teste-t: duas amostras em par para médias para os experimentos Exp 1 e Exp 6 com RR

	Exp 1	Exp 6
Média	07863	0,8174
Variância	0,0555	0,0459
Observações	50	50
Correlação de Pearson	0,9348	
Hipótese da diferença de média	0	
gl	49	
Stat t	-2,4495	
P(T<=t) uni-caudal	0,0089	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0179	
t crítico bi-caudal	2,0095	

Tabela 38 – Teste-t: duas amostras em par para médias para os experimentos Exp 4 e Exp 6 com RR

	Exp 4	Exp 6
Média	0,6684	0,8174
Variância	0,1440	0,04591
Observações	50	50
Correlação de Pearson	0,3522	
Hipótese da diferença de média	0	
gl	49	
Stat t	-2,8930	
P(T<=t) uni-caudal	0,0028	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0056	
t crítico bi-caudal	2,0095	

Tabela 39 – Teste-t: duas amostras em par para médias para os experimentos Exp 3 e Exp 5 com RR

	Exp 3	Exp 5
Média	0,1250	0,0077
Variância	0,0650	0,0306
Observações	50	50
Correlação de Pearson	0,6998	
Hipótese da diferença de média	0	
gl	49	
Stat t	1,8571	
P(T<=t) uni-caudal	0,0346	
t crítico uni-caudal	1,6765	
P(T<=t) bi-caudal	0,0693	
t crítico bi-caudal	2,0095	

Podemos observar na Tabela 38 que o valor para P(T<=t) bi-caudal (0,0056) é inferior ao valor de significância padrão (0,05), o que nos possibilita a concluir que o Experimento 4 é signifitivamente superior ao Experimento 6 mesmo que para a medida MAP isto não se repete, ou seja, o Experimento 4 alcançou 79,01% contra 80,87% do Experimento 6.

Na Tabela 39 apresentamos o resultado do Teste-T entre os experimentos 3 e 5. O Experimento 3 utiliza os três termos com o maior peso dos três documentos melhor classificados julgados como relevantes pelos usuários de acordo com a consulta original. O Experimento 5 utiliza os 5 termos dos três primeiros documentos julgados relevantes pelos usuários de acordo com a consulta original.

Podemos observar na Tabela 39 que o valor para P(T<=t) bi-caudal (0,0693) é superior ao valor de significância padrão (0,05) utilizado neste teste. Isto nos autoriza a concluir que o Experimento 3 não é superior ao Experimento 5 no que que tange ao contexto deste trabalho.

ANEXO A - Regras para a identificação das RLBs

São apresentadas, neste Anexo, as regras para identificação das RLBs, para o Português, utilizadas na ferramenta *RELLEX* (Gonzalez, 2005).

Notação:

AA = adjetivo ou particípio

AJ = adjetivo

AP = particípio

AV = advérbio

CV = conjunto verbal

DT = determinante (artigos definido ou indefinido, ou pronomes demonstrativo ou indefinido)

LD = lado direito

LE = lado esquerdo

PR = preposição

SU = substantivo

VA = verbo auxiliar

VB = verbo

Regras para a identificação de classificações

1. Classificação direta:
   SU_1 SU_2 $\longrightarrow$ =(SU_2, SU_1)
   Condição: há DT antes de SU1, em LD ou LE, sem PR entre DT e SU1.
   Exemplo:
   o goleiro Manga $\longrightarrow$ =(manga,goleiro)

2. Classificação por verbo ”ser”:
   SU_1 ’ser’ SU_2 $\longrightarrow$ =(SU_1, SU_2)
   Condição: não há núcleo no CV e SU_1 é núcleo no LE.
   Exemplo:
   Manga foi goleiro $\longrightarrow$ =(manga,goleiro)

3. Classificação por predicado verbal:
   SU VB $\longrightarrow$ =(SU, h2(VB))
   Condição: SU é núcleo no LE, VB é núcleo no CV e não há preposição ”por” no LD.
   Exemplo:
   cidadão elegeu $\longrightarrow$ =(cidadao,eleitor)

4. Classificação por predicado nominal:
   SU VA AA $\longrightarrow$ =(SU, n_2(AA))
   Condição: SU é núcleo no LE, AA é núcleo no CV e não há preposição ”por” no LD.
   Exemplo:
   animal é rastejante $\longrightarrow$ =(animal,rastejador)

5. Classificação do agente da voz passiva:
   AP 'por' SU $\longrightarrow$ =(SU, n_2(AP))
   Condição: AP é núcleo no CV.
   Exemplo:
   eleito pelo cidadão $\longrightarrow$ =(cidadao,eleitor)

6. Classificação por modificador:
   AA SU ou SU AA $\longrightarrow$ de (AA, SU), se não há n_1(AA) nem n_2(AA)
   Condição: mais próximo SU de AA, em LE ou LD, sem PR entre AA e SU.
   Exemplo:
   biscoito crocante $\longrightarrow$ =(biscoito,crocante)

Regras para identificação de restrições

1. Restrição de objeto por modificador direto:
   AA SU ou SU AA $\longrightarrow$ de (n_1(AA), SU), se há n_1(AA), ou de (SU, n_2(AA)), se há n_2(AA)
   Condição: mais próximo SU de AA, em LE ou LD, sem PR entre AA e SU.
   Exemplos:
   equipe rápida $\longrightarrow$ de(rapidez,equipe) endereço residencial $\longrightarrow$ de(endereco,residencia)

2. Restrição de objeto por modificador preposicionado:
   SU_1 PR SU_2 $\longrightarrow$ PR (SU_1, SU2)
   Condição: mais próximos SU1 e SU2 de PR, sem outra preposição antes deles.
   Exemplo:
   fiscal com experiência $\longrightarrow$ com(fiscal,experiência)

3. Restrição de evento por modificador:
   AV VB ou VB AV $\longrightarrow$ de (n_1(AV), n_1(VB)), se há n1(AV), e de (n_1(VB), n_2(AV)), se há n2(AV)
   Condição: VB é núcleo no CV.
   Exemplos:
   projetou perfeitamente $\longrightarrow$ de(perfeicao,projeto)
   projetou mentalmente $\longrightarrow$ de(projeto,mente)

4. Restrição de modificador por modificador:
   AV AA ou AA AV $\longrightarrow$ de (n_1(AV), n_1(AA)), se há n_1(AV), e de (n_1(AA), n_2(AV)), se há n2(AV)
   Condição: AA é núcleo no CV ou, em LE ou LD, é o mais próximo AA de AV, sem PR entre eles.
   Exemplos:
   adaptado rapidamente $\longrightarrow$ de(rapidez,adaptacao)
   adaptado pessoalmente $\longrightarrow$ de(adaptacao,pessoa)

5. Restrição de objeto por modificador de evento:
   SU VB AV $\longrightarrow$ de (n_1(AV), SU), se há n1(AV), senão de (n_2(AV), SU), se há n2(AV)
   Condição: SU é núcleo em LE, VB é núcleo no CV e não há PR entre VB e AV.
   Exemplos:
   atleta correu facilmente $\longrightarrow$ de(facilidade,atleta)
   feirante construiu artesanalmente $\longrightarrow$ de(artesanato,feirante)

6. Restrição de evento por agente:
   SU VB $\longrightarrow$ por (n_1(VB), SU)
   Condição: SU é núcleo no LE, VB é núcleo no CV e não há preposição ”por” no LD.
   Exemplo:
   forno esquentou $\longrightarrow$ por(esquentamento,forno)

7. Restrição de evento por tema:
   SU VB $\longrightarrow$ de (h1(VB), SU)
   Condição: SU é núcleo no LE, VB é núcleo no CV e não há núcleo no LD.
   Exemplo:
   forno esquentou $\longrightarrow$ de(esquentamento,forno)

8. Restrição de predicado nominal por agente SU VA AA $\longrightarrow$ por (n_1(AA), SU) Condição: SU é núcleo no LE, AA é núcleo no CV e não há preposição ”por” no LD.
   Exemplo:
   prêmio tornou famoso $\longrightarrow$ por(fama,premio)

9. Restrição de predicado nominal por tema:
   SU VA AA $\longrightarrow$ de(n_1(AA), SU)
   Condição: SU é núcleo no LE, AA é núcleo no CV e não há núcleo no LD.
   Exemplo:
   cantor ficou famoso $\longrightarrow$ de(fama,cantor)

10. Restrição de evento por objeto:
    VB SU $\longrightarrow$ de(n_1(VB), SU)
    Condição: SU é núcleo no LD, VB é núcleo no CV.
    Exemplo:
    comprei presente $\longrightarrow$ de(compra,presente)

11. Restrição de predicado nominal por objeto:
    VA AA SU $\longrightarrow$ de (n_1(AA), SU)
    Condição: SU é núcleo no LD, AA é núcleo no CV.
    Exemplo:
    foi comprado o presente $\longrightarrow$ de(compra,presente)

12. Restrição de evento por complemento:
    VB PR SU $\longrightarrow$ PR(n_1(VB), SU)
    Condição: VB é núcleo no CV, PR é primeira preposição no LD e SU é o primeiro substantivo após PR.
    Exemplo: comprei na loja $\longrightarrow$ em(compra,loja)

13. Restrição de predicado nominal por complemento:
    VA AA PR SU $\longrightarrow$ PR(h1(AA), SU)
    Condição: AA é núcleo no CV, PR é primeira preposição no LD e SU é o primeiro substantivo após PR.
    Exemplo:
    ficou calmo sobre a cama $\longrightarrow$ sobre(calma,cama)

14. Restrição de agente por complemento:
    SU_1 VA PR SU_2 $\longrightarrow$ PR(SU_1, SU2)
    Condição: SU_1 é núcleo no LE, não há núcleo no CV, PR é primeira preposição no LD e

SU_2 é o primeiro substantivo após PR.
Exemplo:
equipe está na competição ⟶ em(equipe,competição)

15. Restrição de possuído por possuidor:
SU_1 'ter/possuir' SU_2 ⟶ de (SU_2, SU_1)
Condição: SU1 é núcleo no LE e SU2 é núcleo no LD.
Exemplo:
casa tem porta ⟶ de(porta,casa)

Regras para identificação de associações

1. Associação de agente com tema em evento:
SU_1 VB SU_2 ⟶ n_1(VB) (SU_1, SU_2)
Condição: SU_1 é núcleo no LE, VB é núcleo no CV e SU_2 é núcleo no LD.
Exemplo:
técnico treinou atleta ⟶ treino(tecnico,atleta)

2. Associação de agente com tema na voz passiva:
SU_1 VA AA 'por' SU_2 ⟶ n1(AA) (SU_1, SU_2)
Condição: SU_1 é núcleo no LE, AA é núcleo no CV e SU2 é núcleo no LD.
Exemplo:
atleta foi treinado pelo técnico ⟶ treino(tecnico,atleta)

3. Associação de agente com tema em evento preposicionado:
SU_1 VB PR SU_2 ⟶ n_1(VB).PR (SU_1, SU_2)
Condição: SU_1 é núcleo no LE, VB é núcleo no CV, PR é primeira preposição no LD e SU_2 é o primeiro substantivo após PR.
Exemplo:
turista viajou para a Europa ⟶ viagem.para(turista,europa)

ANEXO B - Diferenças Evidentes

O conceito de evidência, utilizado no cálculo do peso dos descritores, pode ser entendido melhor através dos seguintes exemplos. Considere os dois documentos a seguir, sendo cada um constituído, para simplificar a exemplificação, por uma sentença:

Documento A: "*A fiel governanta, que trabalhou na casa de campo, e o mordomo fugiram*".

Documento B: "*O fiel mordomo, que fugiu para o campo, trabalhou na casa da governanta*".

Considere, também, para que nenhum outro fator influencie o cálculo do peso, que os dois documentos têm comprimentos iguais à média da coleção e que todos os termos têm fator IDF = 1. Com essas condições, na Tabela 40 são apresentados os pesos dos termos lematizados para os dois documentos utilizando a Equação B, baseada em freqüência de ocorrência. Na Tabela 41 são apresentados os pesos dos termos nominalizados e na Tabela D.3, os pesos das RLBs para os dois documentos utilizando as Equações 4.2, 4.3 e 4.4, baseadas em evidência.

Tabela 40 – Peso dos descritores com cálculo baseado em freqüência de ocorrência

	Descritores	doc A ou B/freq	doc A/$W_{t,1}$	doc B/$W_{t,2}$
	campo	1	1	1
	casa	1	1	1
termos	fiel	1	1	1
lematizados	fujir	1	1	1
	governanta	1	1	1
	mordomo	1	1	1
	trabalhar	1	1	1

Considere que, na aplicação da Equação 4.2, são usados os parâmetros k_1 e b com valores 1,2 e 0,75, respectivamente, conforme o que é usualmente adotado.

Tabela 41 – Peso dos termos com cálculo baseado em evidência

	Descritores	doc A/evidência	doc A/$W_{t,A}$	doc B/evidência	doc B/$W_{t,B}$
	campo	0,5	0,65	2,5	1,49
	casa	2,5	1,49	2,5	1,49
termos	fidelidade	1,5	1,22	1,5	1,22
nominalizados	fuga	4,5	1,74	3,5	1,64
	fugitivo	2,5	1,49	1,5	1,22
	governanta	85	1,93	0,5	0,65
	mordomo	3,5	1,64	9,5	1,95
	trabalhador	1,5	1,22	1,5	1,22
	trabalho	3,5	1,64	3,5	1,64

Na Tabela 40 não é possível distinguir termos mais ou menos representativos. Naturalmente, a freqüência de ocorrência restrita a um documento que contém apenas uma sentença pouco

pode contribuir neste sentido. Por outro lado, basta uma sentença para que o cálculo baseado em evidência consiga apontar os descritores mais importantes, conforme pode ser observado na Tabela 41, no caso dos termos, e na Tabela B, no caso das RLBs.

Tabela 42 – Peso das RLBs com cálculo baseado em evidência

	Descritores	doc A/evidência	doc A/$W_{t,A}$	doc B/evidência	doc B/$W_{t,B}$
	=(governanta,fugitivo)	11,0	1,98		
RLBs	=(governanta,trabalhador)	10,0	1,96		
Classificação	=(mordomo,fugitivo)	6,0	1,83	11,0	1,98
	=(mordomo,trabalhador)			11,0	1,98
	de(fidelidade,governanta)	10,0	1,96		
	de(fidelidade,mordomo)			11,0	1,98
	de(fuga,governanta)	13,0	2,01		
	de(fuga,mordomo)	8,0	1,91	13,0	2,01
RLBs	de(trabalho,governanta)	12,0	2,00		
Restrição	de(trabalho,mordomo)			12,0	2,00
	em(trabalho,casa)	6,0	1,83	6,0	1,83
	para(fuga,campo)			6,0	1,83
	por(fuga,governanta)	13,0	2,01		
	por(fuga,mordomo)	8,0	1,91	13,0	2,01
	por(trabalho,governanta)	12,0	2,00		
	por(trabalho,mordomo)			12,0	2,00
	fuga.para(mordomo,campo)			12,0	2,00
RLBs	trabalho.em(governanta,casa)	11,0	1,98		
Associação	trabalho.em(mordomo,casa)			12,0	2,00

Um texto pode ser representado como uma estrutura de dados (Gonzalez & de Lima, 2001). De acordo com o peso baseado em evidência, representações dos documentos A e B na forma de grafos são apresentadas, respectivamente, na Figura B e na Figura B.

Figura 24 – Representação do documento A em grafo

Nesses grafos, os nodos são termos nominalizados e os arcos são RLBs. A espessura das setas e o tamanho dos caracteres são proporcionais aos pesos dos descritores para simular a representatividade dos mesmos.

O termo ”*campo*”, no documento A, e o termo ”*governanta*”, no documento B, não estão presentes em nenhuma RLB porque, de acordo com o modelo TR+, estão envolvidos em relações não evidentes. Essas relações necessitam informações semânticas para serem identificadas. Por exemplo, em ”*trabalhou na casa de janeiro a maio*” a segunda preposição (”*de*”) não associa o que vem depois dela com ”*casa*”, ao contrário de ”*trabalhou na casa de campo*” e de ”*trabalhou na casa da governanta*”. As regras utilizadas para identificar as RLBs não detectam tais diferenças e, assim, não capturam dependências desse tipo.

Nos grafos apresentados ficam visíveis diferenças importantes entre os documentos A e B. Embora eles apresentem os mesmos termos que, por freqüência de ocorrência, não se destacam, a representatividade, com cálculo baseado em evidência, aponta diferenças. Por exemplo, a representatividade do termo ”*governanta*” é grande no documento A e pequena no documento B. Desta forma, uma consulta com o termo ”*governanta*” teria o documento A apontado como mais relevante.

Figura 25 – Representação do documento B em grafo

As RLBs também têm representatividades que mostram diferenças entre os dois documentos. Uma consulta contendo "*fuga de mordomo*" recuperaria os dois documentos, tendo o documento B maior valor de relevância. Já "*fuga de governanta*" recuperaria o documento A como mais relevante.

ANEXO C - Tópicos de Consulta

São apresentados, neste Anexo, os 50 tópicos, para formulação de consultas,utilizados neste trabalho.

- **Tópico 1**
  Título: Abuso sexual.
  Descrição: Recuperar informação sobre abuso sexual sofrido por adulto ou criança.
  Narrativa: Um documento relevante deve relatar ou comentar situação ou situações onde adultos ou crianças foram abusados sexualmente.

- **Tópico 2**
  Título: Acidente rodoviário
  Descrição: Recuperar informação sobre acidente ocorrido em rodovia.
  Narrativa: Um documento relevante deve relatar ou comentar acidente ocorrido em rodovia envolvendo qualquer tipo de dano.

- **Tópico 3**
  Título: Almoço
  Descrição: Recuperar informação sobre almoço.
  Narrativa: Um documento relevante deve relatar ou comentar encontros de pessoas para almoço ou informar sobre pratos servidos em um almoço ou, ainda, sobre preços ou locais deste tipo de refeição.

- **Tópico 4**
  Título: Animação
  Descrição: Recuperar informação sobre animação de pessoas, desenhos ou bonecos.
  Narrativa: Um documento relevante deve relatar ou comentar o ato de alguém se animar ou animar outra pessoa, ou descrever ou comentar a arte de animação de desenhos ou bonecos envolvendo computação gráfica ou qualquer tipo de técnica em produção cinematográfica, de televisão ou alguma mídia digital.

- **Tópico 5**
  Título: Bolsa de valores
  Descrição: Recuperar informação sobre bolsa de valores.
  Narrativa: Um documento relevante deve relatar ou comentar situações que envolvam instituição destinada a operar com ações de companhias ou outros títulos de crédito.

- **Tópico 6**
  Título: Campanha eleitoral de Lula
  Descrição: Recuperar informação sobre a campanha para eleição presidencial de Luis Inácio Lula da Silva.
  Narrativa: Um documento relevante deve relatar ou comentar situações sobre a campanha eleitoral de Luis Inácio Lula da Silva para presidente do Brasil.

- **Tópico 7**
  Título: Caso de cólera
  Descrição: Recuperar informação sobre ações de combate ou efeitos de caso de cólera.
  Narrativa: Um documento relevante deve relatar ou comentar ações de combate ou efeitos de caso (ou casos) de doença infecciosa aguda, contagiosa, que pode manifestar-se sob forma epidêmica, conhecida pelo nome de ”cólera”.

- **Tópico 8**
  Título: Certificação
  Descrição: Recuperar informação sobre certificação.
  Narrativa: Um documento relevante deve relatar ou comentar fatos que envolvam atribuição de algum tipo de certificado a alguém ou a algum produto ou a alguma empresa.

- **Tópico 9**
  Título: Cinema brasileiro
  Descrição: Recuperar informação sobre o cinema brasileiro.
  Narrativa: Um documento relevante deve relatar ou comentar fatos que envolvam o cinema brasileiro, ou seja, filmes produzidos no Brasil com artistas, diretores e recursos nacionais.

- **Tópico 10**
  Título: Cirurgia
  Descrição: Recuperar informação sobre cirurgia médica.
  Narrativa: Um documento relevante deve relatar ou comentar intervenção cirúrgica tanto com objetivo de diagnóstico quanto de cura de alguma doença.

- **Tópico 11**
  Título: Dança
  Descrição: Recuperar informação sobre dança.
  Narrativa: Um documento relevante deve relatar ou comentar fatos relacionados a qualquer tipo de dança, seja clássica, moderna, folclórica ou outro tipo, seja profissional ou realizada por divertimento.

- **Tópico 12**
  Título: Deputado federal
  Descrição: Recuperar informação sobre ações ou características de algum deputado federal.
  Narrativa: Um documento relevante deve relatar ou comentar situação envolvendo algum deputado federal ou descrever características de algum deputado federal.

- **Tópico 13**
  Título: Desemprego
  Descrição: Recuperar informação sobre causas ou efeitos de desemprego.
  Narrativa: Um documento relevante deve relatar ou comentar situações decorrentes de desemprego ou fatos que levam alguém a perder emprego.

- **Tópico 14**
  Título: Digitalização
  Descrição: Recuperar informação sobre o processo de digitalização de documentos.
  Narrativa: Um documento relevante deve descrever ou comentar dispositivos, técnicas ou efeitos de digitalização de textos, imagens ou qualquer outro tipo de documento.

- **Tópico 15**
  Título: Distribuição de renda
  Descrição: Recuperar informação sobre distribuição de renda.
  Narrativa: Um documento relevante deve relatar ou comentar efeitos ou benefícios da distribuição de renda, ou ações destinadas à sua promoção.

- **Tópico 16**
  Título: Drible
  Descrição: Recuperar informação sobre situação em que tenha ocorrido drible.
  Narrativa: Um documento relevante deve descrever ou comentar situação ou efeito de situação em que tenha ocorrido drible, em contexto esportivo ou não, como em ”driblar a concorrência”.

- **Tópico 17**
  Título: Escola de samba
  Descrição: Recuperar informação sobre ações ou características de uma escola de samba.
  Narrativa: Um documento relevante deve descrever características de uma escola de samba, ou relatar ou comentar situação ou evento em que uma escola de samba tenha se envolvido.

- **Tópico 18**
  Título: Exportação
  Descrição: Recuperar informação sobre exportação de algum produto.
  Narrativa: Um documento relevante deve relatar ou comentar fato envolvido com exportação de algum produto.

- **Tópico 19**
  Título: Financiamento agrícola
  Descrição: Recuperar informação sobre financiamento agrícola.
  Narrativa: Um documento relevante deve relatar ou comentar medidas destinadas a promover financiamento agrícola, ou efeitos deste tipo de financiamento.

- **Tópico 20**
  Título: Franquia
  Descrição: Recuperar informação sobre franquia.
  Narrativa: Um documento relevante deve relatar ou comentar fato envolvido com franquia de serviços ou produtos.

- **Tópico 21**
  Título: Globalização
  Descrição: Recuperar informação sobre causas e efeitos de globalização.
  Narrativa: Um documento relevante deve relatar ou comentar causas ou efeitos da globalização, como crescente integração de vários países, em termos de economias, culturas e outros aspectos.

- **Tópico 22**
  Título: Guerra do Golfo
  Descrição: Recuperar informação sobre a Guerra do Golfo.
  Narrativa: Um documento relevante deve relatar aspectos ou comentar causas e conseqüências da Guerra do Golfo.

- **Tópico 23**
  Título: Hotel
  Descrição: Recuperar informação sobre hotel.
  Narrativa: Um documento relevante deve descrever um hotel ou vários hotéis, ou relatar ou comentar preços, promoções, instalações e outros aspectos característico de um ou vários hotéis.

- **Tópico 24**
  Título: Imóvel usado
  Descrição: Recuperar informação sobre imóvel usado.
  Narrativa: Um documento relevante deve descrever características de um ou mais imóveis usados, ou relatar ou comentar transação comercial ou reforma de imóvel usado.

- **Tópico 25**
  Título: Impressora
  Descrição: Recuperar informação sobre impressora.
  Narrativa: Um documento relevante deve descrever características do periférico de computador conhecido como impressora, ou relatar ou comentar fato envolvendo impressora como elemento principal.

- **Tópico 26**
  Título: Informatização
  Descrição: Recuperar informação sobre informatização.
  Narrativa: Um documento relevante deve relatar ou comentar causas, dificuldades, efeitos de informatização de empresa ou serviço, ou descrever alguma informatização realizada.

- **Tópico 27**
  Título: Instrumento musical
  Descrição: Recuperar informação sobre instrumento musical.
  Narrativa: Um documento relevante deve descrever características ou relatar o histórico ou a origem de um instrumento musical, ou explicar a contribuição de um instrumento musical em uma orquestra, banda ou outro tipo de grupo musical.

- **Tópico 28**
  Título: Kit multimídia
  Descrição: Recuperar informação sobre características de um kit multimídia.
  Narrativa: Um documento relevante deve descrever um kit multimídia ou comentar as vantagens de seu uso.

- **Tópico 29**
  Título: Leilão de gado
  Descrição: Recuperar informação sobre a ocorrência e objetivo de um leilão de gado.
  Narrativa: Um documento relevante deve relatar ou comentar a ocorrência de um leilão de gado, ou descrever o tipo de animal leiloado ou transações realizadas.

- **Tópico 30**
  Título: Liderança de campeonato
  Descrição: Recuperar informação sobre liderança de esportista ou equipe em campeonato.
  Narrativa: Um documento relevante deve relatar ou comentar causas ou efeitos da liderança de esportista ou equipe em campeonato que disputa.

- **Tópico 31**
  Título: Medalha de ouro
  Descrição: Recuperar informação sobre disputa ou obtenção de medalha de ouro.
  Narrativa: Um documento relevante deve relatar ou comentar a disputa por medalha de ouro ou a obtenção da mesma.

- **Tópico 32**
  Título: Merenda escolar
  Descrição: Recuperar informação sobre distribuição de merenda escolar.
  Narrativa: Um documento relevante deve relatar ou comentar causas de insucesso, ações para promover ou características da distribuição de merenda escolar, ou apontar responsáveis.

- **Tópico 33**
  Título: Mutuário
  Descrição: Recuperar informação sobre mutuário.
  Narrativa: Um documento relevante deve relatar ou comentar situações envolvendo mutuário, onde este é participante principal.

- **Tópico 34**
  Título: Nudismo
  Descrição: Recuperar informação sobre local ou prática de nudismo.
  Narrativa: Um documento relevante deve relatar ou comentar prática de nudismo, ou descrever características de local desta prática.

- **Tópico 35**
  Título: Passeio de barco
  Descrição: Recuperar informação sobre passeio de barco.
  Narrativa: Um documento relevante deve relatar ou comentar algum passeio onde o percurso tenha sido realizado através de barco.

- **Tópico 36**
  Título: Pastilha de freio
  Descrição: Recuperar informação sobre pastilha de freio.
  Narrativa: Um documento relevante deve descrever vantagens ou desvantagens de algum tipo ou marca de pastilha de freio, ou relatar ou comentar situação onde pastilha de freio tem participação importante.

- **Tópico 37**
  Título: Pintura restaurada
  Descrição: Recuperar informação sobre pintura restaurada.
  Narrativa: Um documento relevante deve descrever vantagens ou desvantagens de algum tipo ou marca de pastilha de freio, ou relatar ou comentar situação onde pastilha de freio tem participação importante.

- **Tópico 38**
  Título: Plano real
  Descrição: Recuperar informação sobre o plano econômico denominado ”Plano Real”.
  Narrativa: Um documento relevante deve relatar ou comentar situações envolvendo o plano econômico conhecido como ”Plano Real”, ou explicar causas e/ou conseqüências de sua implantação.

- **Tópico 39**
  Título: Pólo turístico
  Descrição: Recuperar informação sobre pólo turístico.
  Narrativa: Um documento relevante deve descrever um pólo turístico, ou relatar ou comentar situações características ou localizadas em algum pólo turístico.

- **Tópico 40**
  Título: Produtividade industrial
  Descrição: Recuperar informação sobre produtividade industrial.
  Narrativa: Um documento relevante deve relatar ou comentar ações que trazem aumento ou prejuízo para a produtividade industrial, ou relatar ou comentar efeitos do aumento ou da diminuição da produtividade industrial.

- **Tópico 41**
  Título: Projeto arquitetônico
  Descrição: Recuperar informação sobre projeto arquitetônico.
  Narrativa: Um documento relevante deve descrever aspectos de um projeto arquitetônico, ou comentar sobre os responsáveis, ou relatar ou comentar situações características de um projeto arquitetônico, tanto em relação à sua fase de realização, quanto aos seus efeitos depois de realizado.

- **Tópico 42**
  Título: Propaganda eleitoral gratuita
  Descrição: Recuperar informação sobre propaganda eleitoral gratuita.
  Narrativa: Um documento relevante deve relatar ou comentar situações envolvendo propaganda eleitoral realizada em horário eleitoral gratuito.

- **Tópico 43**
  Título: Publicação eletrônica
  Descrição: Recuperar informação sobre publicação eletrônica.
  Narrativa: Um documento relevante deve descrever o resultado ou características de uma publicação eletrônica ou características de dispositivo ou processo específico para publicação em meio eletrônico.

- **Tópico 44**
  Título: Reajuste salarial
  Descrição: Recuperar informação sobre reajuste salarial.
  Narrativa: Um documento relevante deve relatar ou comentar situações ou campanhas envolvendo tratativas para reajuste de salário, ou comentar reajustes salariais efetivados ou não obtidos.

- **Tópico 45**
  Título: Reciclagem de lixo
  Descrição: Recuperar informação sobre reciclagem de lixo.
  Narrativa: Um documento relevante deve descrever processos para reciclagem de lixo, ou relatar ou comentar medidas para promover a reciclagem de lixo, ou informar sobre responsáveis ou sobre locais onde ocorre ou ocorrerá.

- **Tópico 46**
  Título: Seleção brasileira de futebol
  Descrição: Recuperar informação sobre a seleção brasileira de futebol.

Narrativa: Um documento relevante deve relatar ou comentar situações, disputas ou participantes da seleção brasileira de futebol.

- **Tópico 47**
  Título: Treino oficial
  Descrição: Recuperar informação sobre treino oficial.
  Narrativa: Um documento relevante deve relatar ou comentar situações, participantes ou resultados de um treino oficial de competição automobilística.

- **Tópico 48**
  Título: Uno Mille
  Descrição: Recuperar informação sobre Uno Mille.
  Narrativa: Um documento relevante deve descrever versões do automóvel Uno Mille, ou relatar ou comentar situações onde um veículo dessa marca tem participação importante.

- **Tópico 49**
  Título: Vestibular
  Descrição: Recuperar informação sobre concurso vestibular.
  Narrativa: Um documento relevante deve relatar ou comentar situação característica ou peculiar de um concurso vestibular, de seus participantes ou de seus organizadores.

- **Tópico 50**
  Título: Viagem de carro
  Descrição: Recuperar informação sobre viagem de carro.
  Narrativa: Um documento relevante deve relatar ou comentar situação envolvendo uma viagem realizada através de carro, ou a preparação do mesmo para viagem.

ANEXO D - Documentos julgados como relevantes nos experimentos realizados junto ao Modelo TR+ sem EC

São apresentadas, neste Anexo, as listagens dos documentos julgados relevantes para cada tópico de consulta que consta do Anexo C.

Tópico 1: Abuso sexual

Documentos: 268, 271, 274, 301, 302, 303, 313, 357, 358, 396, 401, 407, 408, 436, 437, 449, 477, 478, 479, 510, 538, 539, 2832

Tópico 2: Acidente rodoviário

Documentos: 139, 260, 307, 308, 410, 452

Tópico 3: Almoço

Documentos: 321, 869, 898, 1006, 1273, 2532, 2872, 3014, 3029, 3044, 3090, 3157, 3181, 3199, 3285, 3313, 3326, 3442, 3648, 3659, 3763, 3786, 3795, 3875, 3879

Tópico 4: Animação

Documentos: 596, 697, 876, 931, 1182, 1226, 1312, 1384, 1401, 1402, 1418, 1498, 1581, 1643, 1646, 1690, 1714, 1910, 1925, 1950, 1956, 1976, 1996, 2006, 2022, 2145, 2152, 2157, 2161, 2169, 2170, 2171, 2172, 2173, 2187, 2199, 2208, 2211, 2212, 2234, 3008, 3021, 3064, 3141, 3472, 3521, 3627, 3737, 3760

Tópico 5: Bolsa de valores

Documentos: 599, 626, 651, 2696, 2790, 2915, 2928, 3771

Tópico 6: Campanha eleitoral de Lula

Documentos: 872, 882, 912, 926, 927, 929, 933, 934, 957, 981, 2275, 2911, 2927, 2931, 2972, 2982, 3244, 3444, 3505

Tópico 7: Caso de cólera

Documentos: 262, 305, 354, 400, 444, 445, 544, 580, 581, 1317

Tópico 8: Certificação

Documentos: 701, 3571

Tópico 9: Cinema brasileiro

Documentos: 1675, 1704, 1717, 1737, 3026, 3035, 3141, 3271, 3956

Tópico 10: Cirurgia

Documentos: 797, 798, 809, 810, 1324, 2842, 2864, 3480

Tópico 11: Dança

Documentos: 275, 464, 1414, 1432, 1436, 1568, 1571, 1572, 1587, 1605, 1606, 1614, 1727, 2616, 2796, 3030, 3205, 3225, 3254, 3272, 3437, 3471, 3684, 3835, 3857, 3858, 3904

Tópico 12: Deputado federal

Documentos: 117, 255, 256, 257, 285, 292, 293, 296, 326, 346, 347, 348, 350, 470, 504, 872, 896, 902, 916, 919, 930, 931, 1662, 2331, 2334, 2335, 2367, 2579, 2581, 3341, 3378, 3429, 3430, 3431, 3432, 3443, 3972

Tópico 13: Desemprego

Documentos: 165, 603, 855, 946, 1713, 2350, 2409, 2690, 2693, 2710, 2903, 2965, 3079, 3852, 3890

Tópico 14: Digitalização

Documentos: 1981, 2023, 2047, 2065, 2079, 2091, 2122, 2125, 2195, 2213, 2225, 2540

Tópico 15: Distribuição de renda
Documentos: 1557, 2236, 2320, 2654, 2961
Tópico 16: Drible
Documentos: 455, 1162, 1197, 1281, 1366, 1374, 1375, 3064, 3474
Tópico 17: Escola de samba
Documentos: 93, 468, 470, 2676, 3295, 3300, 3303, 3317, 3319, 3778, 3959, 4011, 4023
Tópico 18: Exportação
Documentos: 14, 33, 37, 46, 47, 53, 57, 73, 84, 88, 97, 98, 100, 104, 165, 207, 224, 335, 572, 574, 584, 588, 589, 590, 599, 601, 603, 608, 613, 618, 621, 622, 626, 630, 635, 651, 779, 823, 830, 1278, 1635, 1843, 2064, 2073, 2172, 2450, 2513, 2582, 2588, 2591, 2594, 2643, 2725, 2757, 2792, 2814, 2826, 2877, 3175, 3176, 3541, 3637, 3685, 3687, 3756, 3850, 3851, 3975, 4143, 4150
Tópico 19: Financiamento agrícola
Documentos: 1, 6, 14, 29, 30, 87, 88, 91, 97, 101, 104
Tópico 20: Franquia
Documentos: 535, 679, 688, 688, 712, 733, 736, 751, 795, 1125, 1770, 1772, 1773, 1775, 1776, 3518, 3522, 3534, 3548, 3560, 3561, 3562, 3563, 3825
Tópico 21: Globalização
Documentos: 147, 727, 2409, 2584, 2587, 2687
Tópico 22: Guerra do Golfo
Documentos: 130, 132, 844, 2195, 2241, 2855, 2891, 2893, 3160, 3279
Tópico 23: Hotel
Documentos: 466, 595, 614, 894, 1134, 1166, 1167, 1574, 2542, 2740, 2742, 3104, 3262, 3390, 3592, 3594, 3652, 3658, 3663, 3682, 3685, 3687, 3691, 3693, 3694, 3695, 3696, 3701, 3708, 3710, 3713, 3725, 3733, 3752, 3758, 3763, 3764, 3774, 3778, 3779, 3782, 3787, 3793, 3795, 3796, 3814, 3816, 3832, 3842, 3870, 3874, 3875, 3877, 3878, 3879, 3899, 3906, 3909, 3916, 3922
Tópico 24: Imóvel usado
Documentos: 1797, 1798, 1799, 1800, 1801, 1810
Tópico 25: Impressora
Documentos: 1500, 1958, 1964, 1985, 1986, 1993, 2018, 2020, 2021, 2033, 2037, 2039, 2052, 2060, 2061, 2062, 2067, 2083, 2084, 2106, 2107, 2109, 2111, 2120, 2131, 2151, 2160, 2169, 2183, 2205, 2213, 2217, 2219, 3539, 3555
Tópico 26: Informatização
Documentos: 567, 785, 1765, 1794, 1795, 1933, 1936, 2005, 2078, 2650, 3054
Tópico 27: Instrumento musical
Documentos: 1432, 1507, 1515, 1529, 1539, 1586, 1600, 1640, 1727, 2129, 2130, 2208, 2222, 2226, 3131, 3830, 3855
Tópico 28: Kit multimídia
Documentos: 1505, 2050, 2139, 2140, 2209, 2214, 2218, 2219, 2221
Tópico 29: Leilão de gado
Documentos: 8, 24, 28, 39, 60, 62, 63, 65, 67, 79, 106, 111, 112
Tópico 30: Liderança de campeonato
Documentos: 1009, 1010, 1036, 1059, 1060, 1071, 1098, 1099, 1109, 1117, 1170, 1179, 1194, 1195, 1200, 1228, 1241, 1252, 1263, 1266, 1277, 1292, 1293, 1301, 1340, 1350, 1367
Tópico 31: Medalha de ouro
Documentos: 864, 1382, 3371, 3420
Tópico 32: Merenda escolar

Documentos: 273, 542, 1642
Tópico 33: Mutuário
Documentos: 646, 1805, 1822
Tópico 34: Nudismo
Documentos: 3624, 3666
Tópico 35: Passeio de barco
Documentos: 3442, 3625, 3645, 3653, 3655, 3698, 3705, 3711, 3735, 3782, 3832, 3841, 3910
Tópico 36: Pastilha de freio
Documentos: 818, 3293, 4099, 4110
Tópico 37: Pintura restaurada
Documentos: 2264, 2778, 2787
Tópico 38: Plano real
Documentos: 71, 72, 76, 100, 101, 103, 120, 136, 179, 187, 191, 203, 204, 208, 209, 210, 216, 234, 235, 589, 591, 594, 613, 617, 624, 627, 628, 660, 872, 882, 903, 926, 945, 946, 965, 967, 1816, 1831, 1866, 1867, 2379, 2380, 2381, 2409, 2410, 2430, 2557, 2990, 3170, 3175, 3176, 3187, 3194, 3739
Tópico 39: Pólo turístico
Documentos: 3661, 3732, 3757, 3822
Tópico 40: Produtividade industrial
Documentos: 29, 104, 593, 701, 788, 820, 830, 2690, 3567
Tópico 41: Projeto arquitetônico
Documentos: 1558, 1771, 1783, 1832, 1900, 2647, 3757
Tópico 42: Propaganda eleitoral gratuita
Documentos: 876, 885, 889, 890, 892, 915, 928, 931, 932, 933, 962, 963, 979, 981
Tópico 43: Publicação eletrônica
Documentos: 1498, 1505, 1523, 2070, 2071, 2074, 2222, 2224
Tópico 44: Reajuste salarial
Documentos: 129, 130, 133, 134, 168, 174, 177, 178, 191, 195, 214, 216, 218, 219, 222, 226, 248, 250, 600, 646, 734, 752, 770, 946, 2954
Tópico 45: Reciclagem de lixo
Documentos: 289, 555, 571, 572, 1438
Tópico 46: Seleção brasileira de futebol
Documentos: 994, 1005, 1006, 1011, 1024, 1025, 1061, 1131, 1134, 1137, 1138, 1153, 1160, 1161, 1164, 1166, 1167, 1172, 1186, 1189, 1215, 1216, 1242, 1253, 1256, 1257, 1267, 1273, 1286, 1287, 1288, 1289, 1292, 1293, 1337, 1339, 1340, 1342, 1343, 1344, 1347, 1354, 1366, 1367, 1379, 1388, 1392, 1393, 1394, 1395, 1424, 1448, 1452, 1476, 2024, 2230, 2232
Tópico 47: Treino oficial
Documentos: 1036, 1149, 1176, 1177, 1200, 1203, 1206, 1207, 1252, 1311, 3239
Tópico 48: Uno Mille
Documentos: 4046, 4047, 4049, 4054, 4057, 4108, 4127
Tópico 49: Vestibular
Documentos: 154, 512, 1217, 1301, 1313, 1484, 1485, 1487, 1488, 1489, 1493, 1495, 1497, 2512, 3445, 3489, 3671, 3672, 3673, 3674, 3675, 3676, 4031
Tópico 50: Viagem de carro
Documentos: 1332, 1364, 1372, 3783, 3788, 3790, 4046, 4088, 4099, 4100