UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO

EM CIÊNCIA DA COMPUTAÇÃO

“Recuperação de Informação com Auxílio de Extratos Automáticos”

Wilson dos Santos Batista Junior

SÃO CARLOS
Maio/2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

B333ri

Batista Junior, Wilson dos Santos.
Recuperação de informação com auxílio de extratos automáticos / Wilson dos Santos Batista Junior. -- São Carlos : UFSCar, 2006.
130 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Inteligência artificial. 2. Processamento da linguagem natural. 3. Sumarização automática. 4. Sistemas de recuperação da informação. I. Título.

CDD: 006.3 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"Recuperação de Informação com Auxílio de Extratos Automáticos"

WILSON DOS SANTOS BATISTA JÚNIOR

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Membros da Banca:

Profa. Dra. Lúcia Helena Machado Rino
(Orientadora – DC/UFSCar)

Profa. Dra. Maria do Carmo Nicoletti
(DC/UFSCar)

Prof. Dr. Celso Antônio Alves Kaestner
(PUC/PR)

São Carlos
Maio/2006

AGRADECIMENTOS

A Deus e a nosso Senhor Jesus Cristo por não terem me abandonado em nenhum momento de minha vida.

A meus pais, irmão e sobrinha pelo apoio incondicional.

A minha orientadora pela sua contribuição e dedicação a este trabalho.

A Linguateca e ao Jornal Público de Portugal pelo fornecimento da coleção de testes utilizada nos experimentos aqui apresentados.

Aos colegas do LIAA e NILC pelo apoio e contribuição.

Aos meus familiares e aos meus amigos Patrick, Alexandre Gualdi, Thiago Carbonel, Thiago Pardo, Otávio Lemos, Anderson Correia, Seimou, Cristiano Vieira, Caroline Perlin, Gabriela, Carol, Robson, Luciano e Leonardo pela ajuda e força.

A Fábia Regina pelo apoio, carinho, espiritualidade e lição de vida.

A Capes pelo auxílio financeiro que possibilitou a realização deste trabalho.

Aos Padres Pedro e Cesário e a José e Clarice Gualdi pelo apoio e orações.

Resumo

Este trabalho de mestrado investiga a aplicação de técnicas de Sumarização Automática (SA) na Recuperação de Informação (RI), duas áreas que, devido ao crescente número de repositórios de informação digital disponíveis, têm se mostrado muito importantes para a captura de informações.

O intuito do trabalho é verificar a contribuição de extratos gerados automaticamente em duas etapas da RI: a indexação e a realimentação de pseudo-relevantes. O principal objetivo dessas duas etapas é encontrar os termos mais descritivos de um documento. Esse objetivo é relacionado fortemente com o objetivo principal da SA que é condensar as principais partes do documento, o que justifica o emprego.

Para verificar a efetividade das técnicas de SA, foram construídos cinco sistemas que utilizam extratos gerados por um sistema de sumarização que foi considerado útil na tarefa de indicar aos leitores humanos o conteúdo de documentos. Esses sistemas foram avaliados com uma coleção de documentos para testes em RI, escritos em português. Em geral, os resultados apontam que os extratos gerados não foram úteis para a indexação, apresentando desempenho inferior à recuperação que usou o conteúdo completo dos documentos na indexação. No caso da realimentação de pseudo-relevantes, os resultados dos sistemas que utilizam extratos específicos foram superiores aos de um sistema sem esta realimentação em muitos casos. Os melhores resultados foram obtidos pelos extratos específicos multi-documentos, mostrando que esse tipo de extrato pode ser útil para a realimentação de pseudo-relevantes.

Abstract

This dissertation investigates the use of Automatic Summarization (AS) techniques on Information Retrieval (IR), two areas that have attracted growing attention due to the continuous growth of information repositories in digital format.

The main goal of this work is to verify the contribution of extracts generated automatically in two stages of the IR: indexing and Pseudo Relevance Feedback (PRF). In general, the main goal of both phases is to find the most descriptive terms of a given document. This goal in turn is strongly related to that of the AS techniques – to summarize the main parts of a document – which justifies the study.

In order to verify the effectiveness of the AS techniques, we have developed five systems that use extracts generated by a summarizing system that was considered useful in the task of indicating the content of documents to human readers. These systems were assessed through a set of documents written in Portuguese to test IR. In general, the results show that the generated extracts were not useful for indexing, presenting worse performance compared to when using a full document for IR. In the PRF case, however, the results obtained using specific extracts were better than those obtained by a system that does not embed PRF. The best results were obtained when using query-biased multi-documents extracts, indicating that this type of extract may be useful for PRF.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 Introdução

Muita informação, pouco tempo disponível e dificuldade para encontrar uma informação específica são as principais motivações para desenvolvimento e utilização de sistemas de Recuperação de Informação (RI). A grande quantidade de material escrito digital, disponibilizado em bibliotecas digitais, sites jornalísticos, fóruns, correio eletrônico e em outros meios é proporcional à dificuldade em localizar uma informação específica de maneira rápida e precisa. A Recuperação de Informação é a área que há décadas vem direcionando estudos para facilitar o acesso à informação, que pode estar disponível em textos, vídeos, imagens ou hipertextos. Neste trabalho, somente serão tratados textos como itens (ou documentos) a serem recuperados.

O processo de recuperação de documentos que atendem a uma necessidade de informação tem a indexação como um dos seus processos básicos. A indexação visa a construir uma representação sucinta do documento – os termos indexadores – que torne possível e facilite a RI. De forma restrita, um termo indexador é uma palavra-chave (ou um grupo de palavras relacionadas) que possui algum significado próprio. De uma forma mais geral, um termo indexador é simplesmente uma palavra que ocorre no texto de um documento em uma coleção (BAEZA-YATES; RIBEIRO-NETO, 1999). O processo de indexação na RI é, de certa forma, uma tentativa de aproximação da indexação feita por um bibliotecário para facilitar a identificação de um livro.

Outro processo básico da RI é a interação com o usuário. Nessa interação, a comunicação é iniciada, na maioria das vezes, por meio da submissão de consultas formadas pelos usuários com a finalidade de apresentar para o sistema de RI suas necessidades de informação. Após essa submissão, o sistema faz outra interação com o usuário, para apresentar-lhe os documentos recuperados (Apresentação de Resultados), que podem

satisfazer sua necessidade de informação. Essa apresentação deve ser feita de forma que permita ao usuário identificar com facilidade quais documentos apresentados realmente atendem à sua necessidade.

Como forma de direcionar a seleção de novos resultados a serem apresentados, o sistema pode, automática ou interativamente com o usuário, melhorar a consulta apresentada *a priori*. Uma das formas de se fazer isso é a partir da Realimentação de Relevantes (RR). Basicamente, a RR repesa (penalizando ou promovendo), acrescenta ou remove termos de uma consulta baseando-se na ocorrência de termos em documentos considerados relevantes para a consulta.

Pesquisas em RI buscam melhorias nas três etapas citadas acima (indexação, RR e Apresentação dos Resultados). Neste trabalho, será apresentado um estudo da aplicação da Sumarização Automática (SA) nas etapas de indexação e RR. Na indexação, o problema é verificar em que medida sumários podem contribuir para representar documentos de uma coleção, visando a uma recuperação eficiente; na RR, o problema consiste em utilizar o potencial de representação de informações relevantes de um documento para melhorar a consulta do usuário. Em teoria, os sumários agregam essas informações relevantes.

Nas duas etapas, existem trabalhos que empregam técnicas de SA a fim de beneficiar os resultados apresentados por sistemas de RI. Brandow et al. (1995) mostraram que sumários, quando usados como índices, podem aumentar a precisão dos resultados dos sistemas de RI. Acreditando que o sumário de um documento é a melhor fonte para captura de termos importantes para reformular uma consulta, Lam-Adesina e Jones (2001) e Sakai e Sparck-Jones (2001) usaram sumários na RR e mostraram que sumários podem apresentar melhores resultados do que o uso de documentos completos[1].

[1] Um documento completo, neste caso, remete somente ao texto apresentado em um hiperdocumento, por exemplo. Descartamos, neste caso, figuras ou qualquer outra informação não textual. Nesta monografia, portanto, documento e texto são usados indistintamente.

Os sumários utilizados nos trabalhos citados são sumários mono-documento, ou seja, cada sumário é gerado a partir de um único documento. Neste trabalho, o uso de sumários multi-documentos (gerados a partir de vários documentos) também foi considerado. Geralmente, sumários multi-documentos são utilizados para auxiliar o usuário no julgamento de relevância de documentos recuperados, como acontece com o NewsBlaster[2] (MCKEOWN et al., 2002; SCHIFFMAN et al., 2002), ferramenta da University of Columbia.

É possível que os resultados da RI, utilizando técnicas de SA, melhorem ou piorem dependendo da técnica de sumarização utilizada nas duas etapas. Neste trabalho, é apresentada uma variação dos trabalhos anteriores para explorar esse potencial, que consiste principalmente na utilização de um sumarizador que teve sua utilidade comprovada em uma avaliação internacional, a *Document Understanding Conference* 2003. Esse sumarizador é o GistSumm (PARDO, 2005; PARDO et al., 2003; PARDO, 2002), que sumariza um documento de uma forma bem interessante: primeiramente, ele verifica qual é idéia principal do documento. Considera-se que a idéia principal é expressa exatamente por uma sentença do documento, a sentença *gist* (daí o nome *GistSumm,* de ***Gist Summ****arizer*). Depois de selecionada a sentença *gist*, outras sentenças que a complementam são escolhidas para formar o extrato.

Dessa forma, a proposta deste trabalho é verificar a contribuição do GistSumm na RI, especificamente em suas etapas de indexação e RR, possibilitando uma avaliação extrínseca (MANI, 2001) do próprio sumarizador no contexto de RI. Essa avaliação extrínseca é feita de dois modos, ambos possíveis no GistSumm: utilizando-se extratos mono ou multi-documentos. Os primeiros são usados tanto na indexação quanto na RR; os segundos, somente na RR. A utilização de extratos multi-documentos na etapa de RR não foi considerada em nenhum dos trabalhos anteriormente citados.

[2] http://www1.cs.columbia.edu/nlp/newsblaster/, último acesso em 13/06/2006.

Um conjunto de sistemas que utilizam o GistSumm foi construído. Esses sistemas são baseados no modelo vetorial de recuperação (SALTON, 1971), um dos mais clássicos da RI. A coleção escolhida para avaliá-los foi a coleção CHAVE (SANTOS; ROCHA, 2004), que contém artigos jornalísticos escritos em português. A escolha dessa coleção deve-se ao fato de ela ter sido utilizada no CLEF 2004[3] para avaliar sistemas de RI na tarefa de recuperar documentos relevantes a um conjunto de consultas. A coleção CHAVE é a única coleção significativa de documentos em português atualmente disponível, que contém julgamentos de relevância que possibilitam testes de sistemas de RI. Ela foi a primeira coleção em português adotada por um comitê internacional específico da área de RI para avaliação de sistemas. Ao adotá-la, foi possível comparar o desempenho dos sistemas com o desempenho dos sistemas participantes do CLEF 2004 que utilizaram essa coleção.

Neste trabalho, faz-se uma breve introdução sobre a tarefa de RI no Capítulo 2 e os dois capítulos seguintes detalham as etapas de indexação e RR. Em seguida, são descritos os sistemas propostos neste trabalho. No Capítulo 6, são apresentados os resultados obtidos nas avaliações. Por fim, apresentam-se as considerações finais, contribuições, limitações e possíveis extensões deste trabalho.

[3] CLEF é um fórum que tem como objetivo avaliar métodos de RI (http://www.clef-campaign.org, último acesso em 13/06/2006).

2 Recuperação de Informação

A Recuperação de Informação (RI) é a área que estuda métodos para facilitar o acesso a itens de informação. Como já dito anteriormente, esse itens podem ser: textos, vídeos, imagens e hipertextos.

Os sistemas de RI possuem três processos principais: a indexação, a busca e a classificação de documentos, com o objetivo de satisfazer às necessidades de informação de seus usuários, expressas por consultas. De acordo Salton (1968), todo sistema de RI pode ser descrito como sendo formado por um conjunto de documentos, um conjunto de consultas, e um mecanismo para determinar quais documentos atendem às consultas que geralmente são disparadas por uma interação com o usuário e, a partir de seus termos de busca, o mecanismo começa a processar a coleção, verificando quais documentos podem ser relevantes. Esses documentos passam por um processo de ordenação, baseado em algum método que indique seus graus de relevância.

Documentos e consultas são geralmente representados por vetores contendo informação sobre as ocorrências dos termos da coleção nos documentos e nas consultas. A etapa que constrói representações para o documento é chamada de indexação (discutida no Capítulo 3).

Os sistemas de RI tentam selecionar as representações de documentos que atendem à necessidade de informação representada pela consulta. Assim, uma representação ruim pode prejudicar todo o processo de recuperação. Depois de identificados os documentos que poderão atender à necessidade de informação, eles são apresentados como saída do sistema. Os usuários poderão, então, visualizar os documentos ou seus descritores (por exemplo, os títulos dos documentos) a fim de verificar se esses documentos são realmente pertinentes. É possível também que os usuários indiquem quais dos documentos apresentados são realmente

relevantes e assim o sistema poderá fazer uma outra recuperação com base nessas informações; esse processo é chamado de Realimentação de Relevantes (discutido no Capítulo 4).

No restante deste capítulo os modelos mais tradicionais de recuperação de documentos são apresentados juntamente com algumas abordagens de avaliação de desempenho dos sistemas de RI.

2.1 Modelos de sistemas de recuperação

A determinação de relevância de um documento, ou seja, a determinação de que um documento atende a uma consulta, depende do modelo de recuperação empregado. Os modelos mais tradicionais são: o booleano, o probabilístico (ROBERTSON; SPARCK-JONES, 1976) e o vetorial (SALTON, 1971). No modelo booleano (ou lógico), a consulta é representada por termos concatenados por operadores lógicos (AND, OR, NOT) resultando em uma expressão lógica. Os termos podem ser palavras ou frases. Os documentos recuperados são aqueles que contêm os termos que satisfazem à expressão lógica representada pela consulta. Em geral, considera-se o *matching* exato, resultando em uma classificação booleana: ou as representações dos documentos satisfazem à expressão lógica e, portanto, são recuperados, ou são descartados.

A principal desvantagem desse modelo é que o usuário deve saber especificar muito bem sua consulta, ou seja, ele deve conhecer a linguagem lógica para formalizá-la. Outra desvantagem é que o modelo não possui mecanismos para determinar os graus de relevância dos documentos. Dessa forma, não é possível ordená-los no momento de apresentá-los como resultados da busca.

O modelo probabilístico recupera documentos através de probabilidades de relevância. A relação – probabilidade de o documento ser relevante à consulta dividida pela probabilidade

de ele não ser relevante – é o indicador de relevância desse documento (BAEZA-YATES; RIBEIRO-NETO, 1999). Essa relação permitirá a atribuição de graus de relevância para os documentos recuperados, possibilitando sua ordenação, o que não é possível com o modelo booleano.

O modelo probabilístico, criado por Robertson e Sparck-Jones (1976), assume que os termos são independentes e cada documento é representado por um vetor binário em que cada elemento indica a ausência ou presença de um termo da coleção no documento. Essa técnica é chamada de *Binary Independence Retrieval*. Com os termos independentes, a probabilidade de um documento ser relevante ou irrelevante é determinada pela probabilidade de seus termos pertencerem a documentos relevantes ou irrelevantes, respectivamente.

No modelo vetorial, os documentos e a consulta são representados por vetores de termos ponderados. O peso determina a importância de um termo para descrever o conteúdo de um documento; utilizar as freqüências do termo é uma forma para determinar esse peso (algumas medidas de atribuição de pesos são detalhadas no Capítulo 3). Os documentos são recuperados de acordo com a similaridade do vetor que os representa com aquele que representa a consulta.

Como exemplo de cálculo de similaridade considere-se uma coleção de documentos contendo apenas dois termos: ‘bagre’ e ‘pintado’. Tanto os vetores que representam os documentos quanto o que representa a consulta terão dimensão igual a 2. A primeira posição do vetor indica o peso do termo ‘bagre’ no documento ou na consulta, e a segunda posição indica o peso do termo ‘pintado’. Seja a coleção formada por três documentos: D1, D2 e D3, e os vetores que os representam <0.1,0.3>, <0.4,0.1>, <0.8,0.6>[4]. Considere-se também uma consulta (Q), cujo vetor que a representa seja <0.8, 0.8>. Os documentos são recuperados de acordo com a proximidade dos seus vetores em relação ao vetor de Q no espaço vetorial; no

[4] Nesta monografia o ponto (.) será utilizado ao invés da vírgula (,) como separador decimal.

exemplo em questão D3 certamente seria selecionado. A similaridade entre uma consulta Q e um documento D quaisquer pode ser calculada por meio do produto interno entre os vetores:

$$sim_{pi}(D,Q) = \sum_{i=1}^{t} (t_{iD} * t_{iQ}) \quad (1)$$

em que t_{iD} representa o peso do termo i no documento D, t_{iQ} representa o peso do termo i na consulta Q e t é igual ao número de termos indexadores da coleção de documentos. O produto interno irá medir o que há de comum entre os dois vetores. Os valores de similaridade são ordenados e, então, os documentos com maior grau de similaridade com a consulta são recuperados. No nosso exemplo, D1, D2 e D3 teriam os graus de similaridade iguais a 0.32, 0.4 e 1.12, respectivamente.

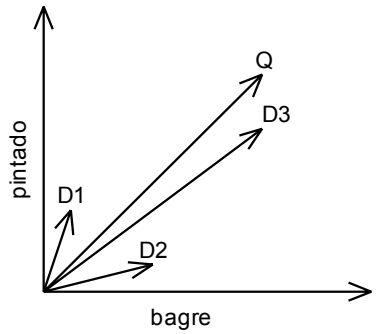


Figura 1 – Três documentos e uma consulta representados em um espaço bidimensional

Além do produto interno, outras funções como o coeficiente de Dice são utilizadas para o cálculo de similaridade entre documentos e consultas no modelo vetorial. A função para o cálculo de similaridade usando o coeficiente de Dice é mostrada a seguir (SALTON; McGILL, 1983).

$$sim_{dic}(D,Q) = \frac{2\sum_{i=1}^{t} (t_{iD} * t_{iQ})}{\sum_{i=1}^{t} t_{iD}^{2} + \sum_{i=1}^{t} t_{iQ}^{2}} \quad (2)$$

Ao contrário do produto interno, o coeficiente de Dice é normalizado entre 0 e 1. A normalização procura minimizar a influência dos comprimentos dos documentos em sua recuperação. Isso é importante quando medidas como a freqüência do termo no documento são utilizadas na ponderação dos termos, já que, nesses casos, documentos longos tendem a ter termos com pesos mais altos que documentos curtos.

Há casos em que o coeficiente de Dice consegue um melhor desempenho que o produto interno, pelo fato de considerar as características individuais da consulta e do documento. Por exemplo, num caso em que a consulta é representada por <1,1>, um documento Dx é representado por <0.5,0> e Dy é representado por <0.3,0.2>; a similaridade de Dy com a consulta, usando o produto interno, é igual à similaridade de Dx com a consulta, embora Dy apresente os dois termos da consulta com pesos maiores que zero e, para a consulta, os dois termos são igualmente importantes. O coeficiente de Dice já apresentará uma similaridade maior para Dy, devido ao fato de o denominador do coeficiente de Dice permitir distinguir a contribuição de cada termo para a consulta (pelos quadrados individuais de cada peso).

Outro coeficiente que distingue as contribuições de cada termo para a consulta é o cosseno de similaridade (SALTON, 1968), apresentado na Eq. 3, que mede a similaridade de um documento com a consulta pelo cosseno do ângulo entre os vetores que os representam. Neste caso o ângulo é utilizado para verificar a proximidade entre os vetores. Quanto maior o ângulo, menor será sua similaridade; e quanto menor, mais o cosseno do ângulo aproxima-se de 1, significando que os dois vetores são muito similares. Quando os vetores são totalmente distintos, por exemplo, quando os termos com pesos positivos para a consulta possuem peso nulo no documento, o ângulo será de 90º e o cosseno será nulo (SALTON, 1968), indicando que o documento não é relevante para a consulta.

$$sim_{cos}(D,Q) = \frac{\sum_{i=1}^{t}(t_{iD} * t_{iQ})}{\sqrt{\sum_{i=1}^{t} t_{iD}^{2}} * \sqrt{\sum_{i=1}^{t} t_{iQ}^{2}}} \quad (3)$$

Para o exemplo da Figura 2, pode se notar que o ângulo α (entre D2 e Q) é maior que o ângulo β (entre D1 e Q). Logo, o cosseno de similaridade indicará uma maior similaridade entre o documento D1 e a consulta Q do que entre o documento D2 e a consulta Q.

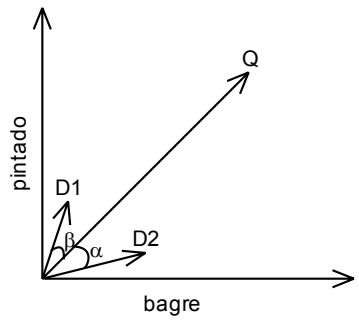


Figura 2 – Ângulos entre as representações vetoriais dos documentos D1 e D2 e a consulta Q

A diferença entre as medidas do cosseno (sim_{cos}) e do coeficiente de Dice (sim_{dic}) está no fato da segunda ser mais sensível aos pesos dos termos atribuídos ao documento que o cosseno, pelo fato de o denominador de Dice ter a tendência de produzir maiores variações que o denominador do cosseno. Não se sabe se essa diferença é uma vantagem ou desvantagem já que não foram encontrados resultados comparativos entre as duas medidas.

2.2 Avaliação de sistemas de Recuperação de Informação

Basicamente, existem duas características a serem avaliadas em um sistema de RI: efetividade e eficiência. Segundo Salton e McGill (1983), a efetividade indica a habilidade de o sistema satisfazer a necessidade de informação, recuperando documentos relevantes e descartando documentos irrelevantes. Existem várias medidas de efetividade utilizadas para comparar se uma técnica de recuperação é melhor do que outra. Algumas dessas medidas

também consideram um critério de satisfação do usuário: a ordem de apresentação dos documentos recuperados.

A eficiência de recuperação simplesmente mensura o custo e o tempo necessários para executar as tarefas de recuperação e pode ser verificada pela complexidade dos algoritmos de busca ou por medidas do tempo de indexação, do tempo de resposta ou do consumo de memória. Um sistema que leva muito tempo para indexar e retornar os documentos ou consome muito espaço de memória do computador é considerado pouco eficiente.

A avaliação de efetividade pode ser feita de duas formas distintas: com ou sem a interação com usuários. Por ser menos custosa e mais padronizada, a avaliação sem a interação com usuários tem sido tradicionalmente utilizada em diversas pesquisas e também nas avaliações conjuntas como CLEF e TREC[5].

Nas avaliações dos sistemas desenvolvidos durante este trabalho, foi considerada somente a medida de efetividade, não sendo o foco imediato a implantação dos sistemas em um ambiente de busca real. Desse modo, não foram feitas quaisquer considerações sobre sua eficiência. Devido a essa limitação do trabalho ao contexto de satisfação das necessidades de informação, são apresentados neste capítulo somente alguns dos principais métodos de avaliação de efetividade de sistemas de RI.

A subseção 2.2.1 apresenta as coleções de testes para avaliar os sistemas de RI sem a interação com o usuário. Na subseção 2.2.2, as medidas utilizadas nas avaliações de efetividade são apresentadas.

2.2.1 Coleções de testes

As coleções de testes de avaliações de sistemas de RI são formadas por um conjunto de documentos, um de tópicos e um de informações indicando julgamentos de relevância

[5] *Text REtrieval Conference* (TREC) tem como objetivo avaliar métodos de RI para encorajar pesquisas em coleções de textos de larga escala. (http://trec.nist.gov, último acesso em 13/06/2006).

previamente estabelecidos, ou conjunto de referência. As coleções são usadas da seguinte forma: o sistema recupera os documentos da coleção que podem ser relevantes a cada um dos tópicos da coleção e, a seguir, o conjunto de referência é utilizado para verificar quais dos documentos retornados são realmente relevantes. Nas próximas subseções, os três conjuntos e sua utilização serão detalhados.

2.2.1.1 Conjunto de documentos

Os documentos de coleções de testes mais tradicionais como os da TRECs, CLEFs, CACM ou ISI[6], são, em sua maioria, artigos jornalísticos e científicos caracterizados por serem textos curtos e com poucos tópicos, fato que, de certa forma, facilita a tarefa dos sistemas em recuperá-los.

A coleção CACM, por exemplo, é formada por 3204 artigos sobre Ciência da Computação publicados na *Communications of the ACM* entre os anos de 1958 e 1979. Os documentos dessa coleção, além dos textos principais, possuem outras informações como: nome do autor, título, data de publicação, sumários, categorias às quais os documentos pertenciam e co-citações (BAEZA-YATES; RIBEIRO-NETO, 1999). Já a coleção ISI é formada por 1460 artigos sobre Ciência da Informação. Além dos textos, seus documentos contêm informações como: nome do autor, título, sumário, co-citações. Essas duas coleções são consideradas pequenas, quando comparadas com as coleções de bibliotecas digitais, portais de notícias e as coleções disponíveis na Web, como um todo. Segundo Cormack et al. (1998), coleções com um milhão de documentos ou mais são necessárias para a avaliação de sistemas de RI modernos. Com o objetivo de construir coleções com maior quantidade de documentos e fazer comparações entre vários sistemas e técnicas, foi criada a TREC, cujas coleções possuem uma quantidade de documentos bem mais expressiva que as outras.

[6] As coleções CACM e a ISI podem ser obtidas em http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/, último acesso 13/06/06.

A TREC faz parte de uma iniciativa do National Institute of Standards Technology (NIST) e do Departamento de Defesa norte-americano para fornecer suporte para avaliações de métodos de RI em larga escala. Desde 1992, anualmente é promovida uma nova edição da TREC, que fornece uma nova coleção de testes. Os documentos das sucessivas TRECs são formados basicamente por artigos jornalísticos provenientes do *Wall Street Journal, San Jose Mercury News, Finantial Times, LA Times*, entre outros. Além do texto, os documentos possuem informações como data de publicação e autoria.

O esforço de criar grandes coleções como as fornecidas pela TREC não se resume simplesmente em coletar material e pedir autorização aos autores para uso dos seus documentos. O problema maior consiste em registrar o julgamento humano de relevância, pois em coleções pequenas (como a CACM) é possível efetuá-lo, porém é quase impossível em coleções muito grandes.

2.2.1.2 Conjunto de tópicos

Os tópicos de uma coleção de testes indicam as necessidades de informações de um suposto usuário e simulam intenções de busca. Um tópico é construído artificialmente como uma descrição não ambígua da necessidade de informação. Ele é geralmente dividido em três campos: título, descrição e narrativa. O título é o principal descritor do tópico, a descrição é uma sentença especificando a necessidade de informação, enquanto que a narrativa é um relato mais detalhado daquilo que um documento relevante deve ou não conter. Um exemplo de tópico é mostrado na Figura 3.

<top>
<num> Number:033
<title> Topic: Impact of foreign textile imports on U.S. textile industry
<desc> Description: Document must report on how the important of foreign textiles or textile products has influenced or impacted on the U.S. textile industry.
<narr> Narrative: The impact can be positive or negative or qualitative. It may include the expansion or shrinkage of markets or manufacturing volume or an influence on the methods or strategies of the U.S. textile industry. “Textile industry” includes the production or purchase of raw materials; basic processing techniques such as dyeing, spinning, knitting, or weaving; the manufacture and marketing of finished goods; and also research in the textile field.
</top>

Figura 3 – Tópico 033 da TREC (extraído de Tombros e Sanderson, 1998)

De acordo com Voorhees (1998), os assessores da TREC responsáveis pelos julgamentos de relevância criam os tópicos e, baseando-se numa estimativa do número de documentos relevantes e no balanceamento dos temas dos tópicos, selecionam alguns deles. Uma vez construído o conjunto de tópicos, consultas podem ser derivadas automática ou manualmente, para a recuperação de documentos da coleção correspondente. Segundo Buckley e Voorhees (2000), o uso de diferentes consultas afeta o comportamento da recuperação, pois algumas representações de um tópico são melhores que outras.

Além de permitir a derivação de consultas, os tópicos também são usados na avaliação dos sistemas. Segundo Voorhees (1998), na TREC o assessor que cria o tópico é quem verifica se o conjunto de resultados automáticos atende à necessidade de informação descrita. Dessa forma, o que se avalia é a relevância do documento ao tópico e não à consulta formulada.

O número de tópicos da coleção é uma questão importante. Segundo Buckley e Voorhees (2000), uma coleção de testes deve ter um número razoável de tópicos para que os resultados dos experimentos tenham boa confiabilidade. A TREC, por exemplo, tem usado 25 tópicos no mínimo e 50 como norma.

2.2.1.3 Conjunto de referência

Julgar a relevância de um documento é verificar se ele atende à necessidade de informação representada por um tópico. Essa, indubitavelmente, é a parte mais polêmica da coleção de testes devido ao caráter subjetivo do julgamento da relevância de documentos, fator determinante da variação de escolha entre os juízes.

A tarefa de analisar a relevância de vários documentos a um tópico é um tanto quanto árdua, porém possível quando a coleção possui poucos documentos. Segundo Sanderson e Joho (2004), as coleções das décadas de 60, 70 e parte da década de 80 (como por exemplo, Cranfield e CACM) eram pequenas; a Cranfield e a CACM tinham 1400 e 3204 documentos, respectivamente. Dessa forma, era possível verificar a relevância dos documentos através de uma análise exaustiva da coleção. No entanto, a criação das grandes coleções tornou impossível a tarefa de analisar documentos um a um.

A fim de poder utilizar um grande número de documentos em avaliações sem a necessidade dessa análise, surgiu a idéia de *pooling*, proposta por Sparck-Jones e Van Rijsbergen (1975), que consiste numa estratégia para efetuar julgamentos de relevância de maneira não exaustiva, já que somente um conjunto de documentos recuperados por um conjunto de sistemas é analisado. Os documentos que não estiverem nesse conjunto são considerados irrelevantes sem passar por qualquer julgamento. O método de *pooling* tem sido padrão para criar coleções de testes como as utilizadas na TREC e no CLEF (SANDERSON; JOHO, 2004).

Nesse método, os assessores examinam os primeiros k documentos com maiores similaridades com as consultas de cada um dos conjuntos de resultados retornados pelos n sistemas. Quanto maiores k e n, mais representativo será o conjunto de referência. Porém, isso

aumenta a quantidade de documentos que devem ser julgados já que ela é proporcional a *k* e *n*. Por exemplo, para uma das tarefas da TREC 6 o valor de *k* foi igual a 100 e o de *n* igual a 30, necessitando de aproximadamente 60 mil julgamentos para 50 tópicos (COMARCK et al., 1998).

Como é possível perceber, o método de *pooling* é extremamente dependente dos resultados dos sistemas de RI utilizados. Caso os sistemas tenham uma recuperação apresentando um número baixo de documentos relevantes da coleção, independente do número de *k* e de *n*, o julgamento de relevância será muito distante do real, ou seja, do julgamento exaustivo. Esse julgamento de relevância poderá trazer problemas com o reuso da coleção de testes. Segundo Braschler e Peters (2004), a questão de reuso de coleções construídas através de técnica de *pooling* pode ser crítica, pois um *pool* incompleto (ou seja, distante do real) pode colocar os experimentos que não contribuíram para o *pool* em desvantagem. De fato, esses experimentos podem recuperar documentos relevantes que não foram julgados sendo, portanto, classificados como irrelevantes.

Como na TREC e no CLEF o método de *pooling* é utilizado nos julgamentos de relevância, podemos dizer que há a possibilidade de esses julgamentos serem incompletos, pois pode haver documentos relevantes não apontados como tal. Apesar disso, o *pooling* possibilitou a viabilização de avaliações de grande escala. Outro ponto a favor dessa metodologia é que ela apresenta como relevantes somente os documentos que os sistemas avaliados conseguiram recuperar.

2.2.1.3.1 Classificações de relevância

Um documento é julgado relevante na TREC se algum trecho seu é relevante, independentemente de quão pequeno esse trecho seja com relação ao restante do documento[7]. Caso isso não ocorra, ele é considerado irrelevante. Esse tipo de julgamento é binário e seu uso tem sido criticado pela falta de realismo (SORMUNEN, 2002), já que geralmente usuários de sistemas de RI julgam a relevância dos documentos em escalas subjetivas de relevância.

Como o julgamento binário não permite distinguir graus variados de relevância, pesquisadores propõem coleções com vários níveis de relevância. Sormunen (2002), por exemplo, apresenta um experimento com vários níveis de julgamento em uma nova avaliação da relevância dos documentos do *pooling* da TREC 7 e TREC 8 referentes a 38 tópicos. Para isso, ele estabelece quatro níveis de relevância:

(0) Irrelevante - o documento não contém qualquer informação sobre o tópico;

(1) Superficialmente relevante - há pouca informação relevante ao tópico (geralmente, uma sentença). Nesse caso, o documento não contém outras informações além das presentes na descrição do tópico;

(2) Relevante - o documento contém mais informações que a descrição do tópico, mas a apresentação não é exaustiva. Em casos de documentos abordando vários tópicos, somente alguns sub-temas os cobrem, geralmente, um parágrafo ou duas a três sentenças sobre o tópico;

(3) Altamente relevante - o documento discute o tema do tópico exaustivamente. No caso de documentos abordando vários tópicos, a maioria dos sub-temas abordam assuntos a eles referentes, geralmente, diversos parágrafos ou pelo menos quatro sentenças sobre o tópico.

[7] http://trec.nist.gov/data/reljudge_eng.html, acessado em 13/06/2006.

Os assessores de Sormunen classificaram os documentos das coleções com esses níveis de relevância. Após a classificação, esses níveis de relevância foram comparados com os julgamentos originais realizados pelos juízes da TREC. O total de documentos julgados foi de 5737, dos quais 48% foram julgados relevantes pelos assessores da TREC. Segundo o julgamento de Sormunen (2002), 61% dos documentos foram considerados irrelevantes, 20% superficialmente relevantes, 13% relevantes e 6% altamente relevantes.

Ao comparar seus julgamentos com os efetuados na TREC, Sormunen constatou que 25% dos documentos considerados relevantes na TREC foram julgados irrelevantes no seu experimento, 36% dos documentos considerados relevantes na TREC foram julgados superficialmente relevantes e o número de documentos julgados relevantes, mas que foram considerados irrelevantes na TREC, foi pequeno (cerca de 1%). Porém isso não quer dizer que somente 1% dos documentos relevantes da coleção foi desconsiderado na TREC, pois somente o *pool* de documentos foi analisado.

Como se nota, os julgamentos podem variar bastante e uma parte significante dos documentos considerados relevantes na TREC tinha relevância superficial. Essa última constatação influi fortemente nas perspectivas dos sistemas que consideram somente extratos genéricos dos documentos como índices, como poderá ser visto neste trabalho.

2.2.2 Métricas de avaliação de sistemas de RI

A RI tem uma tradição bem estabelecida em executar experimentos com coleções de testes para comparar a efetividade de diferentes abordagens de recuperação. Segundo Buckley e Voorhees (2000), a natureza do experimento muitas vezes serve para especificar os critérios de avaliação a serem utilizados para determinar se uma abordagem é melhor que a outra. Como o comportamento dos sistemas de recuperação é suficientemente complexo, diferentes

medidas de efetividade já foram propostas. Segundo Saracevic (1995), as medidas para avaliar as diversas abordagens devem ser escolhidas de acordo com os objetivos dos sistemas e com sua habilidade em recuperar documentos relevantes. Devido a esse foco, elas não consideram a eficiência nem a forma de apresentação dos documentos recuperados para o usuário. TREC e CLEF seguem esse mesmo foco, considerando as medidas tradicionais de precisão e revocação, propostas inicialmente por Kent et al. (1955).

A precisão é definida como a proporção do material (documentos) recuperado que é relevante, enquanto a revocação é a proporção do material relevante da coleção que é recuperado (SALTON; McGill, 1983). Basicamente, a revocação mede a habilidade de um sistema em recuperar todos os documentos relevantes presentes na coleção e a precisão mede a habilidade de um sistema em recuperar somente documentos relevantes. As fórmulas para calcular precisão e revocação são apresentadas a seguir.

$$\text{precisão} = \frac{\text{nroRelevantesRecuperados}}{\text{nroTotalRecuperados}} \tag{4}$$

$$\text{revocação} = \frac{\text{nroRelevantesRecuperados}}{\text{nroTotalRelevantes}} \tag{5}$$

em que nroRelevantesRecuperados é igual ao número de documentos relevantes recuperados, nroTotalRecuperados é o número de documentos recuperados e nroTotalRelevantes é o número de documentos relevantes da coleção.

Diferentemente das medidas precisão e revocação, que ressaltam a efetividade de um sistema em recuperar documentos relevantes, a medida *fallout* ressalta sua habilidade em ignorar e, portanto, impedir a recuperação de documentos irrelevantes. Ela é calculada pela fórmula apresentada na Eq. 6.

$$\text{fallout} = \frac{\text{nroIrrelevantesRecuperados}}{\text{nroTotalIrrelevantes}} \tag{6}$$

em que nroIrrelevantesRecuperados é o número de documentos irrelevantes recuperados e nroTotalIrrelevantes é o número de documentos irrelevantes da coleção. Uma alta medida de *fallout* indica que o sistema recupera muitos documentos irrelevantes da coleção e, portanto, seu desempenho não é satisfatório. A *fallout* é pouca usada nas avaliações de RI pelo fato da atenção maior ser dada para as medidas de precisão e revocação.

Como os sistemas de RI geralmente retornam os documentos em ordem decrescente de similaridade com a consulta, e os usuários tendem a verificar os documentos na ordem de apresentação, é necessário considerar essa ordem na avaliação dos sistemas. Segundo White et al. (2002), a ordem de apresentação é fundamental, pois os usuários restringem-se a consultar a primeira página dos resultados apresentados.

Uma das medidas que considera a ordem de apresentação dos documentos é a precisão para os primeiros *k* documentos recuperados (P_k), calculada somente para esse subconjunto:

$$P_k = \frac{\text{nroRelevantesRecuperadosk}}{\text{k}} \tag{7}$$

em que nroRelevantesRecuperadosk é o número de documentos relevantes dentre os *k* documentos recuperados. Dessa forma, P_k é a proporção de documentos relevantes dentre os primeiros k documentos recuperados.

Partindo dessa idéia, a R-Precision consiste em um refinamento de P_k: ela é a precisão para os primeiros k documentos recuperados, em que k é o número de documentos relevantes para o tópico da coleção. Diferentemente de P_k, ela indica a capacidade do sistema em recuperar todos os documentos relevantes para o tópico da coleção nas primeiras posições da lista de documentos recuperados.

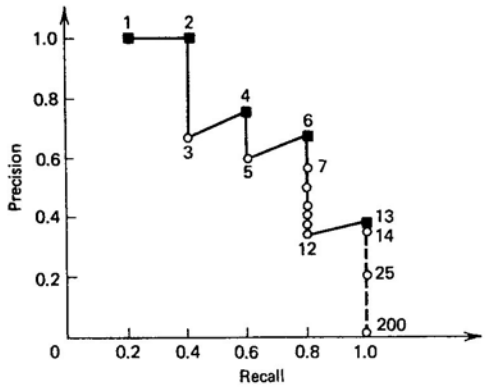
O valor de precisão para os primeiros k documentos recuperados pode ser visualizado com a revocação alcançada quando k documentos são recuperados. Assim, verifica-se a dependência entre a medida de precisão e a porcentagem dos documentos relevantes da

coleção recuperados pela relação entre precisão e revocação. O fato é que, à medida que o número de documentos recuperados aumenta, a revocação tende a aumentar. Assim, questiona-se se a medida de precisão de um sistema pode apresentar valores que indiquem uma boa proporção de documentos relevantes à medida que a revocação aumenta.

O número de graus distintos de revocação é igual ao número de documentos relevantes da coleção para o tópico. Por exemplo, se para um tópico há cinco documentos relevantes na coleção, então há cinco graus distintos de revocação: (1/5), (2/5), (3/5), (4/5), (5/5). A Figura 4 (SALTON; McGILL, 1983, p. 166) apresenta uma tabela (a) mostrando a variação de precisão e revocação conforme o número de documentos recuperados aumenta. Nessa tabela, n representa o número de documentos recuperados (chamado anteriormente de k), *Recall* é a revocação alcançada quando n documentos são recuperados, e *Precision* é a precisão para os primeiros n documentos recuperados (chamada anteriormente de P_k). Dados os conjuntos de pares de precisão e revocação da tabela da Figura 4, o gráfico apresentado em (b) é construído.

Recall-precision after retrieval of n documents			
n	Document number (x = relevant)	Recall	Precision
1	588 x	0.2	1.0
2	589 x	0.4	1.0
3	576	0.4	0.67
4	590 x	0.6	0.75
5	986	0.6	0.60
6	592 x	0.8	0.67
7	984	0.8	0.57
8	988	0.8	0.50
9	578	0.8	0.44
10	985	0.8	0.40
11	103	0.8	0.36
12	591	0.8	0.33
13	772 x	1.0	0.38
14	990	1.0	0.36

(a)



(b)

Figure 5-2 Display of recall and precision results for a sample query. (Collection consists of 200 documents in aerodynamics.) (a) Output ranking of documents in decreasing query-document similarity order and computation of recall and precision values for a single query. (b) Graph of precision versus recall for sample query of Fig. 5-2a.

Figura 4 – Exemplo de resultados de revocação e precisão para uma consulta

Na tabela, o 'x' ao lado do número do documento (segunda coluna) indica que o documento é relevante, sua ausência indica que o documento é irrelevante. Como se pode observar, a cada documento irrelevante, a revocação se mantém estável, enquanto a precisão diminui. E a cada documento relevante encontrado, tanto a precisão quanto a revocação aumentam[8]. Como os sistemas tendem a recuperar um número maior de documentos relevantes dentre os primeiros recuperados, conforme o número de documentos considerados aumenta, a precisão diminui e a revocação aumenta. O fato é que, quanto mais documentos são considerados nos cálculos, maior é o número de documentos irrelevantes dentro desse conjunto.

[8] No caso de precisão igual a um, ela se mantém estável.

O gráfico mostra que, quando a revocação é igual a 0.2, só há um valor de precisão; porém, quando um valor maior de revocação é considerado, além de a precisão diminuir, são verificados vários valores de precisão para o mesmo grau de revocação (por exemplo, quando a revocação é igual a 0.8). O fato de haver vários valores de precisão para um único grau de revocação mostra claramente que, quando o grau de revocação é alto, a precisão tende a diminuir. Um sistema com uma recuperação perfeita obteria a precisão igual a 1 para todos os graus de revocação e somente quando o grau de revocação fosse igual a 1 o gráfico apresentaria queda nos seus valores de precisão, pois, como todos os documentos relevantes da coleção já foram recuperados, cada documento, considerado a partir desse ponto, seria irrelevante.

Segundo Salton e McGill (1983) gráficos de precisão por revocação, como o da Figura 4**,** têm sido criticados pelo número de parâmetros obscuros. Por exemplo, o tamanho do conjunto de documentos recuperados e o tamanho da coleção não podem ser obtidos através do gráfico. Além disso, problemas podem ser encontrados quando um gráfico contínuo é produzido a partir de pontos discretos, isto é, o valor de precisão é conhecido exatamente quando a revocação é igual a 0.2, porém quando a revocação é igual a 0.4 não são claros os vários valores de precisão. Outro problema ocorre quando é necessário construir um gráfico para verificar o desempenho do sistema para um conjunto de tópicos, pois alguma estratégia tem que ser adotada para calcular a média das precisões, já que pode haver tópicos que apresentam mais de um valor de precisão por grau de revocação.

Para resolver esse último problema faz-se uma interpolação dos valores de precisão para os graus de revocação obtidos para cada tópico. O gráfico da Figura 5 (SALTON; McGILL, 1983, p. 167) mostra a curva formada pelos valores de precisão interpolados para um único tópico. A curva do gráfico é obtida desenhando-se um segmento de linha horizontal de um grau de revocação para outro, da direita para a esquerda, ou seja, do maior grau de

revocação para o menor. O primeiro ponto do segmento é o ponto de maior precisão. Após esse processo, a Figura 5 exibe um único valor de precisão para cada grau de revocação. A precisão interpolada é considerada, então, a precisão máxima atingida quando um determinado grau de revocação é alcançado.

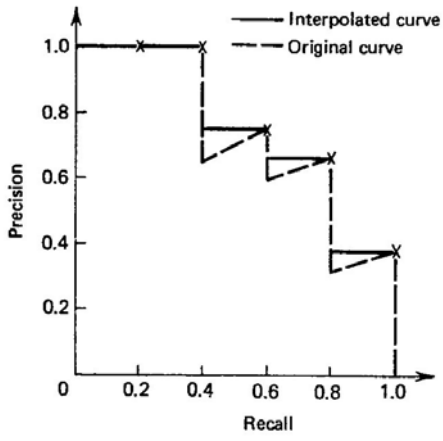


Figure 5-3 Interpolated recall-precision curve for sample query of Fig. 5-2. (Ranks of relevant items are 1, 2, 4, 6, 13.)

Figura 5 – Curva precisão por revocação interpolada

Geralmente, para calcular a precisão média interpolada, são somente consideradas as precisões interpoladas para 11 graus padrões de revocação: 0.0, 0.1, ..., 1.0. Em gráficos de precisão por revocação interpolada, os sistemas com ótimo desempenho tendem a apresentar valores de precisão próximos a 1 para todos os 11 graus padrões de revocação, porém a maioria apresenta o desempenho típico em que a precisão é alta quando o grau de revocação é baixo e vice-versa. Com o gráfico de precisões interpoladas, é possível verificar o declive médio da precisão conforme os valores de revocação aumentam.

Os 11 graus padrões de revocação, ou a curva da precisão por revocação formada por eles, podem ser utilizados para comparar dois sistemas. Segundo Raghavan et al.(1989), tradicionalmente um sistema A é considerado melhor que um sistema B, se, em todo grau de

revocação, a precisão do sistema A é maior que a do sistema B. Se esta superioridade não for mantida para todos os graus de revocação, uma média dos valores de precisão para cada grau de revocação apresentados pelos sistemas é calculada e comparada.

Com os valores máximos de precisão não interpolada para graus de revocação distintos, a MAP (*Mean Average Precision*) que, segundo Voorhees (1998), é considerada o sumário estatístico do sistema, pode ser calculada. A MAP é a média das precisões obtidas quando cada documento relevante da coleção é encontrado na lista de documentos recuperados. Essa média é calculada para cada tópico e depois uma média desses valores resultantes é calculada. No caso de uma recuperação perfeita, em que todos os documentos relevantes são apresentados nas primeiras posições da lista de recuperados, a MAP é igual a 1.

Diante de tantas medidas fica difícil saber qual utilizar para análise e comparação de desempenho de sistemas de RI. Nesse caso, os objetivos do sistema sob avaliação, além dos critérios de satisfação dos usuários, podem ajudar a defini-las. Por exemplo, se o interesse é atender a usuários que verificam somente os primeiros documentos recuperados, então P5 pode ser uma boa opção. Agora, se a opção é atender a usuários que verificam somente os primeiros resultados, mas o número de documentos recuperados que eles analisam é variável, a MAP seria uma boa escolha.

Todas as medidas apresentadas aqui, com exceção da MAP, são calculadas de forma individual para cada tópico e o desempenho do sistema é avaliado com o valor médio obtido para todos os tópicos da coleção. Alguns pesquisadores (por exemplo, Hull, 1993) acreditam que é necessário realizar testes estatísticos de significância que visem a analisar as diferenças de resultados de dois sistemas para cada tópico. O objetivo é verificar se um sistema não obteve média superior ao outro somente porque teve um desempenho melhor para um número pequeno de tópicos.

Van Rijsbergen (1979) argumenta que o uso de testes estatísticos é impróprio para a RI, pois os valores de precisões e revocações são discretos, enquanto a maioria dos testes estatísticos utilizados é baseada na suposição de que a população analisada tem distribuição contínua. É fácil notar que as medidas apresentadas aqui são discretas, pois todas variam de acordo com o número de documentos relevantes da coleção. Por exemplo, se só há um documento relevante na coleção, a precisão para os primeiros cinco documentos recuperados só poderá ser igual a 0.0 ou 0.2. e a revocação só poderá ser igual a 0.0 ou 1.0. De acordo com Hull, se um número grande de tópicos e de documentos relevantes for considerado, os valores discretos pode aproximar-se de uma distribuição contínua.

O teste dos sinais, segundo Van Rijsbergen, é um dos mais recomendáveis, pois não tem muita dependência da natureza ou forma da população envolvida, ignorando completamente a magnitude das diferenças. Se um método de recuperação tem melhor performance que outro mais freqüentemente que uma média esperada, então existe uma forte evidência de que o método é superior (HULL, 1993).

Diferentemente do teste dos sinais, o teste t emparelhado[9] considera a magnitude das diferenças entre os métodos. Segundo Hull, o teste t assume que o erro segue uma distribuição normal, mas, às vezes, trabalha bem quando ela não é seguida. Se a diferença média entre os métodos for grande, comparada com seu erro padrão, então o método será significativamente diferente.

Para fazer os dois testes, isto é, o dos sinais e o teste t, é necessário definir a probabilidade de rejeitar a hipótese nula (H_0) quando verdadeira. H_0 é a hipótese de que os resultados dos dois sistemas são estatisticamente equivalentes, ou seja, a diferença não é estatisticamente significante. A probabilidade de rejeitar H_0 quando ela é verdadeira é

[9] Se uma amostra tem alguma relação com a outra, diz-se que elas são dependentes. Tais amostras costumam ser chamadas emparelhadas, ou ligadas (TRIOLA, 1999). No caso da recuperação, as amostras são resultados para os tópicos e o resultado do sistema A para o tópico q tem relação com o resultado do sistema B para o mesmo tópico.

chamada de nível de significância (α). Segundo Triola (1999), o valor de α é tipicamente predeterminado e são comuns as escolhas de α=0.05 e α=0.01. Se a hipótese nula é rejeitada, os dados indicam evidências de que os dois sistemas não são equivalentes. Segundo Hull, se um teste de significância aponta que a diferença entre os métodos não é significante, isso não necessariamente significa a inexistência de diferença entre os métodos, mas que o teste foi incapaz de detectá-las.

3 Indexação de Documentos

Indexação é o processo de se associar uma estrutura de índice a um conjunto de dados. A estrutura de índice é utilizada para aumentar o desempenho na obtenção do resultado de uma consulta (ELMASRI; NAVATHE, 1998). Em sistemas de RI, ela contém, por exemplo, os termos mais representativos de documentos, que possibilitam a recuperação daqueles potencialmente relevantes a uma consulta.

Segundo Salton e McGill (1983), a tarefa de indexação consiste, primeiro, na seleção de termos capazes de representar o conteúdo do documento e, segundo, da atribuição a cada termo de um peso ou valor, refletindo sua presumida importância para o propósito de identificação de conteúdo.

Os termos usados na indexação (chamados termos indexadores) podem ser todas as palavras do documento ou somente algumas palavras-chave. Em vez das palavras ou palavras-chaves, seus radicais[10] podem ser utilizados. Nesse caso, as palavras com os mesmos radicais são tratadas como um único elemento. As ferramentas para a geração de radicais de palavras serão detalhadas na próxima seção.

Belew (2000) cita que um índice é uma relação de mapeamento de um documento em um conjunto de palavras-chave. Como uma busca é geralmente feita a partir das palavras-chave, é estabelecido o mapeamento inverso, isto é, o mapeamento das palavras-chave em documentos, dando origem ao arquivo de índice invertido, que será detalhado na Seção 3.2.

A identificação das palavras-chave (ou termos-chave) não é uma tarefa fácil. De acordo com Salton e McGill, entre todas as operações necessárias em RI, a mais crucial e provavelmente a mais difícil consiste em atribuir termos apropriados e identificadores capazes

[10] Radicais são gerados por um processo chamado *stemming* que retira os sufixos de uma palavra para gerar o respectivo radical.

de representar o conteúdo dos documentos. Se o índice for ruim, todo o processo de recuperação estará seriamente comprometido.

A polissemia (uma palavra com vários significados) pode afetar o processo de indexação, conseqüentemente o de recuperação. O problema está no fato de que uma palavra polissêmica pode ser termo indexador para um conjunto de documentos que apresentam diferentes conceitos. Por exemplo, uma busca pelo termo 'bancos' pode recuperar tanto documentos sobre estabelecimentos ou sociedades mercantis de crédito quanto documentos sobre assentos.

Salton e McGill citam que uma das fontes para achar os termos indexadores é o próprio texto dos documentos, seus títulos ou seus sumários manuais. A freqüência do termo tem sido um indício muito utilizado para a determinação de termos indexadores. Segundo Luhn (1958), o uso da freqüência das palavras no texto é baseado no fato de que um escritor normalmente repete certas palavras conforme ele avança ou varia seus argumentos. Palavras com freqüências altas ou baixas não seriam bons termos indexadores, mas palavras com freqüência medianas, sim (SALTON; McGILL, 1983). As palavras com freqüências altíssimas são geralmente artigos, conjunções, interjeições e preposições e compõem as chamadas *stoplists*, que são utilizadas para identificar palavras que não são boas discriminadoras de conteúdo.

Um termo indexador pode ser formado por várias palavras (termo composto) ou por uma única palavra (termo simples). Um termo composto possui mais informação que um termo simples, por exemplo, se um documento é indexado pelo termo simples 'Fernando', então ele pode conter informações sobre uma ou várias pessoas chamadas 'Fernando', enquanto que o termo composto 'Fernando Henrique Cardoso' é mais específico, indicando que o documento pode conter informações sobre o ex-presidente da república do Brasil.

A indexação por termos simples é mais fácil, pois a extração de termos pode ser feita usando apenas a tarefa de tokenização. A grande dificuldade está em extrair termos compostos, já que é necessário identificar o relacionamento entre as palavras. De acordo com Fagan (1987), as estratégias para identificar termos compostos variam de simples procedimentos baseados em freqüências e coocorrências de palavras até sofisticados métodos que empregam a análise sintática. Fagan apresenta um experimento usando termos compostos como indexadores, obtidos por um processo simples. Para que um conjunto de palavras coocorrentes fosse selecionado como um termo composto, cinco parâmetros foram pré-estabelecidos: o número máximo de elementos de um termo composto, a janela de ocorrência das palavras (uma sentença), a proximidade entre as palavras (palavras adjacentes excluindo stopwords), a freqüência mínima para a palavra principal do termo composto, e a freqüência do termo composto nos documentos da coleção.

Fagan mostra que a utilização de termos compostos na indexação produz ganhos de efetividade, quando comparada ao uso de termos simples para cinco coleções (CACM, INSPEC, Cranfield, MEDLINE, CISI). Ele também mostra que a identificação de termos compostos sem a utilização de análise sintática gera muitos erros e pode prejudicar a efetividade da recuperação. Ou seja, ele conclui que a estrutura sintática ajuda a reconhecer melhor o relacionamento entre as palavras que irão formar o termo composto. Por exemplo, para o sintagma nominal: '*parallel and sequential algorithms'*, o método não sintático poderia gerar o termo composto: '*parallel and sequential'* sem considerar o núcleo do sintagma nominal, enquanto que, com análise sintática, é possível verificar que '*sequential*' e '*parallel*' são modificadores de '*algorithms*'. Dessa forma, gerar-se-iam os termos compostos '*parallel algorithms*' e '*sequential algorithms'*.

Na próxima seção será apresentado o processo que faz a extração dos radicias de uma palavra, seguida da explicação da construção de uma estrutura de índice. Na Seção 3.3, alguns

métodos de ponderação de termos são mostrados, seguida da seção que apresenta trabalhos que usam sumários na indexação de documentos.

3.1 Stemming

Como já dito anteriormente, em vez de usar palavras ou palavras-chave na indexação, seus radicais podem ser utilizados. O processo para extração de radicais de uma palavra é chamado *stemming*, esse processo é feito por sistemas chamados de *stemmers*. Os *stemmers* fazem o agrupamento de variantes morfológicas de uma palavra em uma única classe que será representada por um radical. Plurais, formas de gerúndio e sufixos de tempo e pessoa são exemplos de variações sintáticas que podem ser agrupadas em uma classe de equivalentes. Quando se usa *stemming* na indexação, todos os termos agrupados em uma classe são tratados como um único termo indexador. Segundo Orengo e Huyck (2001), o *stemming* é largamente utilizado em procedimentos de processamento de texto em RI, baseando-se na suposição de que uma consulta que contém a variante de uma palavra implica o interesse por documentos que contêm outras variantes da mesma palavra. Um outro objetivo do uso de *stemming* é diminuir a dimensão do arquivo de índices e contribuir para o processo do cômputo das similaridades entre documentos e consultas, aumentando a revocação do sistema de RI. Assim, o stemming é visto como um dispositivo para aumento de revocação, pelo fato de promover a expansão das consultas originais com formas de palavras relacionadas. Às vezes ele pode também contribuir para a melhoria da precisão em uma específica taxa de revocação, promovendo o aumento do ranque dos documentos relevantes.

Existem dois principais problemas que um *stemmer* pode apresentar: *overstemming* e *understemming*. O *overstemming* se dá quando a cadeia de caracteres removida não é um sufixo, mas parte do radical. O *understemming* ocorre quando um sufixo não é removido

completamente. O primeiro pode fazer com que diferentes termos que não sejam variantes de uma mesma palavra sejam agrupados em uma mesma classe. Dessa forma, documentos que contêm esses diferentes termos podem ser recuperados. No segundo caso, as variantes de uma palavra não seriam agrupadas em uma mesma classe. Dessa forma, a busca por todas as variantes de uma palavra não seria completa.

Há diversos tipos de *stemming*, dentre eles: os baseados em listas de regras de remoção de sufixos (por exemplo, PORTER, 1980 e LOVINS, 1968); os baseados em dicionários (por exemplo, FULLER; ZOBEL, 1998); e os baseados em córpus (por exemplo, XU; CROFT, 1998). Essa diversidade parece não estar presente nos *stemmers* para a língua portuguesa: os mais conhecidos são baseados em listas de regras de remoção de sufixos. São eles: Snowball[11], o de Caldas Jr. et al. (2001), o de Orengo e Huyck (2001) e o Pegastemming[12]. O stemmer de Caldas Jr. et al. é baseado no *stemmer* de Porter (1980).

De maneira geral, o *stemming* para a língua portuguesa geralmente retira de uma palavra desinências nominais e verbais, sufixos e vogais temáticas. Desinências nominais indicam gênero e número dos nomes e as desinências verbais indicam tempo, modo, pessoa e número nos verbos. Sufixos são elementos mórficos adicionados a fim de formar novas palavras ou mudar seu sentido (por exemplo, '-dade', como sufixo de lealdade). Já a vogal temática agrega-se ao radical de uma palavra para receber as desinências.

O *stemmer* de Orengo e Huyck (2001) foi construído de acordo com as características do português, diferentemente do *stemmer* para o português baseado no algoritmo de Porter (1980). Cabe ressaltar que a morfologia da língua portuguesa é muito mais complexa que a morfologia da língua inglesa. Isso pode ser constatado através das diferentes formas verbais, superlativos, gênero, etc. Essa complexidade torna mais difícil a construção de *stemmers* para o português do que para o inglês.

[11] http://snowball.tartarus.org/algorithms/portuguese/stemmer.html, último acesso em 13/06/2006.
[12] Pegastemming é de autoria de Marco Gonzalez (PUC-RS), disponível em http://www.inf.pucrs.br/~gonzalez/ri/pesqdiss/analise.htm, último acesso em 13/06/2006.

Em Orengo e Huyck (2001), os radicais produzidos para 2800 palavras foram testados com a finalidade de verificar erros do *stemming* e corrigi-los. Ao final desse processo, o *stemming* tinha 98% de precisão. Quando testado para um conjunto de 1.000 palavras diferentes das iniciais, a precisão foi de 96%.

Orengo e Huyck compararam seu *stemmer* com outro também para o português, o Muscat baseado no *stemmer* de Porter. Este foi obtido em um site (http://open.muscat.com), que atualmente encontra-se indisponível. Numa avaliação utilizando 1.000 palavras, o *stemmer* de Orengo e Huyck produziu 96% dos radicais corretamente, enquanto o Muscat produziu 71%. Após outra série de avaliações, eles verificaram que seu *stemmer* apresenta menos problemas de *understemming* e *overstemming* que o Muscat, concluindo que era melhor que a versão do *stemmer* de Porter para o português.

Chaves (2003) comparou o *stemmer* de Orengo e Huyck com o Pegastemming. Para essa análise, foram considerados os radicais gerados pelos dois *stemmers* para 500 palavras. As palavras foram divididas nas seguintes categorias gramaticais: substantivo, verbo, adjetivo, pronome, contração de preposição e advérbio. Para todas as categorias, o *stemmer* de Orengo e Huyck apresentou melhor desempenho. As taxas de *understemming* e *overstemming* de Orengo e Huyck também foram menores, exceto a de *overstemming* para substantivos.

3.2 Arquivo de índices invertido

A função do arquivo de índices produzido no processo de indexação é possibilitar uma busca por documentos efetiva e eficiente. A busca é, em muitos sistemas, disparada quando se realiza a interação com os usuários. Portanto, é o momento em que o tempo gasto pelo sistema deve ser minimizado.

Segundo Navarro (1999), um arquivo de índice invertido é composto por dois elementos: a lista de termos indexadores e a lista de ocorrências, sendo estas as referências aos documentos indexados pelos termos indexadores. Um segmento ilustrativo de arquivo invertido é mostrado na Figura 6 (SALTON; McGILL, 1983), em que a palavra 'uva', por exemplo, ocorre nos documentos 3, 7, 9 e 11. A lista de termos indexadores é formada por uma lista de palavras que poderiam ter sido canonizadas ou passadas por um processo de *stemming*. A lista de ocorrências é formada simplesmente por identificadores do documento, o 'id doc' na figura. Outras informações poderiam ser incluídas, tais como: número total de ocorrências e as posições dos termos no documento. O número total de ocorrências seria necessário para calcular algumas medidas (próxima seção); armazenar as posições das palavras seria importante na busca por termos adjacentes.

vocabulário	Ocorrências (id doc)
banana	4,6,8
laranja	2,3,4,5,6
maçã	1,3,5,7
uva	3,7,9,11

Figura 6 – Segmento ilustrativo de um arquivo de índices invertido

Uma busca por documentos com os termos 'laranja' e 'uva' usando *matching* exato (ou seja, 'laranja AND uva') retornaria o conjunto de documentos indexados por ambos os termos, resultando somente no documento '3'.

Como há palavras que ocorrem em milhares dos documentos da coleção, fazendo com que o espaço para armazenamento seja muito grande, Scholer et al. (2002) apresentam técnicas importantes para a redução deste espaço, que compactam a lista de ocorrências formadas pelos identificadores dos documentos juntamente com a freqüência da palavra no documento.

Na próxima seção serão apresentadas algumas medidas usadas para ponderar os termos indexadores de acordo com suas supostas importâncias.

3.3 Ponderação dos termos indexadores

Como já foi dito, na indexação baseada no modelo vetorial, são atribuídos aos termos indexadores pesos que irão refletir a importância do termo na identificação de conteúdo do documento. Uma das maneiras mais simples de atribuição de pesos é a feita de acordo com a presença ou ausência do termo; neste caso, o peso 1 indica a presença do termo e o 0 indica a sua ausência. Assim, o documento seria recuperado de acordo com a presença dos termos da consulta.

Em vez dessa atribuição de pesos binários, são empregadas outras medidas de ponderação que facilitam o ranqueamento dos documentos de acordo com as funções de similaridade entre consulta e documento. Uma dessas medidas é a idf – *inverse document frequency* (Sparck-Jones, 1972). A idf mede a raridade de um termo na coleção, pois atribui maiores pesos aos termos que ocorrem em poucos documentos da coleção indexada. Segundo Sparck-Jones, esses termos raros têm grande poder de discriminação dos documentos de uma coleção.

A medida idf pode ser representada pela seguinte equação:

$$idf_k = \log(N/n_k) \qquad (8)$$

em que N é o número de documentos da coleção, e n_k é o número de documentos da coleção que contêm o termo k. A medida idf é uma medida global, ou seja, o peso do termo k indica seu poder discriminatório na coleção inteira, independentemente dos documentos a que ele pertença. Inicialmente, o logaritmo era especificado para a base 2, porém a sua base não é, em geral, importante.

Uma outra medida muito utilizada para ponderação de termos é a tf-idf (SALTON; McGILL, 1983). A tf-idf, além de considerar a medida idf, considera a medida tf, que é

baseada na ocorrência do termo k em um documento específico. A fórmula a seguir é usada para calcular a medida tf-idf do termo k no documento i.

$$\text{tf-idf}_{k,i} = \text{freq}_{k,i} * \log(N/n_k) \quad (9)$$

sendo $\text{freq}_{k,i}$ a freqüência do termo k no documento i. A medida tf-idf apresentará maiores valores para termos que são raros na coleção e também que tenham freqüência alta no documento i. Ela é baseada no fato de que, se um termo ocorre muito em um documento, mas pouco na coleção, ele pode ser um bom discriminador daquele documento.

As medidas de ponderação que utilizam diretamente a freqüência dos termos nos documentos podem fazer com que documentos mais longos tenham um valor de similaridade com as consultas muito maior do que os documentos mais curtos, pois estes últimos possivelmente têm uma freqüência menor de termos. Para tentar evitar que documentos curtos sejam prejudicados no processo de atribuição de graus de similaridades, estratégias de normalização dos valores de freqüência podem ser adotadas. Uma delas é dividir a freqüência do termo pela freqüência do termo de maior ocorrência no documento. Isso fará com que o peso atribuído ao termo varie de 0 a 1.

Essa freqüência normalizada pode modificar a medida tf-idf, formando a tf-idf com tf normalizada pela freqüência máxima. A tf-idf com tf normalizada é definida através da Eq. 10.

$$\text{tf-idf}_{k,i} = \text{freq}_{k,i} * \log(N/n_k) / \max\{\text{freq}_{l,i}\} \quad (10)$$

em que $\max\{\text{freq}_{l,i}\}$ é a freqüência do termo com maior ocorrência no documento i.

Além de ser normalizada pela freqüência máxima de um documento, a freqüência pode ser normalizada pelo número total de termos no documento. Uma terceira forma de normalizar a freqüência de termos é usar o logaritmo natural da freqüência mais uma constante, técnica chamada de logaritmo da freqüência do termo. O logaritmo, neste caso,

reduz o efeito das grandes variações entre as freqüências dos termos de um documento (LEE, 1995). A idéia é que não é porque um termo ocorre quatro vezes mais que outro que será quatro vezes mais importante que este. A medida, usando a constante igual a 1, é mostrada na Eq. 11, em que $lnfreq_{k,j}$ é o logaritmo da freqüência do termo k no documento j.

$$lnfreq_{k,j} = \ln(tf_{k,j} + 1) \quad (11)$$

Llopis et al. (2004) aplicam esse logaritmo duas vezes na freqüência de ocorrência do termo para reduzir ainda mais os efeitos das variações das freqüências. A equação resultante é mostrada abaixo:

$$lnlnfreq_{k,j} = 1 + \ln(1 + \ln(tf_{k,j} + 1)) \quad (12)$$

sendo $lnlnfreq_{k,j}$ o peso do termo k no documento j.

Existem outras medidas que são utilizadas na ponderação de termos. As apresentadas aqui são algumas das principais usadas em sistemas baseados no modelo vetorial. Nos experimentos realizados durante este trabalho, somente as medidas representadas pelas equações 10 e 12 foram consideradas.

3.4 Sumários na indexação

Como já mencionado, a opção para a determinação automática de termos indexadores de documentos é considerar todas as palavras contidas nele. Outra opção seria utilizar sumários como índices de documentos, como propõem Sakai e Sparck-Jones (2001) e Brandow et al. (1995). O termo sumário é utilizado aqui para representar tantos os sumários produzidos por humanos (sumários manuais) quanto os formados automaticamente por um processo de extração de sentenças do texto-fonte (extratos). Sakai e Sparck-Jones utilizaram em seus experimentos esses dois tipos de sumários, enquanto que Brandow et al. utilizaram somente extratos.

Existem processos de Sumarização Automática, mais especificamente os de extração, que são bem semelhantes a alguns processos de indexação. Essa semelhança é indicada por Salton (1963): as palavras mais significativas são usadas como termos indexadores para caracterizar os documentos e as sentenças mais significativas, que contêm um grande número de palavras significativas, são usadas como sumários dos documentos. Dessa forma, pode-se pensar que, se um sumário é composto por um grande número de palavras significativas, então ele pode ser uma boa fonte para a seleção de termos indexadores.

Além de conter palavras significativas, a outra motivação para o uso de sumários como índices é que eles podem reduzir a dimensão do arquivo de índices e, assim, possibilitar o aumento da eficiência de busca. A redução da dimensão ocorre porque a lista de ocorrências dos termos indexadores é reduzida quando os sumários são considerados.

Sakai e Sparck-Jones (2001) analisaram se um sistema de RI que emprega sumários genéricos como índices poderia produzir resultados semelhantes aos gerados por um sistema de RI que utiliza índices construídos a partir de documentos. Sumários genéricos são os construídos com os possíveis tópicos principais dos documentos, em oposição aos sumários específicos, que são aqueles construídos baseados em uma consulta do usuário.

Segundo Sakai e Sparck-Jones, sua escolha por sumários na indexação foi baseada na idéia de que um bom sumário genérico deveria reter o conteúdo principal do documento original, descartando os segmentos periféricos, sendo assim um bom elemento para indicar se o documento é altamente relevante a uma consulta. ‘Documento altamente relevante’ é o termo utilizado por eles para indicar um documento da TREC indicado como relevante para um tópico e que também possuía a maior parte do seu conteúdo referindo-se ao tópico em questão.

Para testar sua hipótese, Sakai e Sparck-Jones utilizam parte de uma coleção de testes provenientes das TREC 1, 2 e 3, cujos documentos contêm, além dos textos, sumários

manuais em um campo do documento chamado SUMMARY. Essa sub-coleção totaliza 38.884 documentos. Como nem todos os documentos da coleção foram selecionados, eles escolheram somente os tópicos para os quais havia pelo menos cinco documentos relevantes, dentre os 38.884, constituindo 30 tópicos. Para 25 deles, pelo menos três documentos foram considerados altamente relevantes.

Além dos sumários manuais, foram testados, na tarefa de indexação, extratos com diferentes taxas de compressão[13] (95%, 90%, 70% e 50%). Foram também gerados extratos de mesmo tamanho dos sumários manuais. Em todos os casos, os sumários continham no mínimo, uma sentença. Além disso, o título do documento foi incluído nos extratos, independentemente da taxa de compressão, para melhorar o processo de busca, sabidamente, o título de autoria indica a idéia principal do documento correspondente.

Três tipos de extratos foram utilizados:

- *lead*: o extrato *lead* contém sentenças iniciais do documento;
- tfidf: o extrato foi formado por sentenças que possuíam maior pontuação obtida através do somatório da pontuação de cada termo da sentença. Os termos do documento passaram pelo processo de remoção de *stopwords* e *stemming*; então, atribuíam-se pontuações aos termos restantes de acordo com a medida tf-idf;
- 1p_tfidf: esse extrato foi composto por sentenças do primeiro parágrafo do documento e, adicionalmente, foram incluídas sentenças de maior pontuação (pontuação tf-idf). Elas foram incluídas enquanto a taxa de compressão não era excedida.

Os experimentos com esses extratos foram feitos utilizando-se um sistema de busca baseado no modelo probabilístico. As consultas foram geradas artificialmente utilizando-se os

[13] A taxa de compressão é a proporção do conteúdo do documento que foi reduzido no sumário.

campos *title* e *description* dos tópicos da TREC (veja Capítulo 2 para ilustração sobre os tópicos da TREC).

A avaliação com todos os documentos relevantes da TREC mostrou que o uso de extratos não produziu resultados semelhantes aos obtidos com o uso dos documentos completos. As tabelas 1 e 2, respectivamente, apresentam a precisão média e a precisão para os primeiros 10 documentos recuperados (P10) quando os extratos com taxas de compressão iguais a 95%, 90%, AL, 70% e 50% foram utilizados na indexação. A taxa de compressão AL representa os extratos de tamanho proporcional ao tamanho dos sumários manuais.

Tabela 1 – Precisão média com extratos para indexação, todos os documentos relevantes

Extrato	95%	90%	AL	70%	50%
lead	0.101	0.116	0.132	0.172	0.172
tfidf	0.121	0.138	0.169	0.196	0.219
1p_tfidf	0.105	0.130	0.165	0.197	0.219

A recuperação que usa o documento apresentou uma precisão média de 0.25 e uma P10 de 0.36. Como pode ser visto, esses valores foram superiores aos alcançados pelos extratos (tabelas 1 e 2). A precisão média obtida com indexação por extratos *lead* com 50% de taxa de compressão foi a que mais se aproximou da precisão média obtida com os documentos completos. Para P10, o melhor resultado foi obtido com extratos com 70% de taxa de compressão.

Tabela 2 – P10 com extratos para indexação, todos os documentos relevantes

Extrato	95%	90%	AL	70%	50%
lead	0.276	0.296	0.308	0.320	0.320
tfidf	0.280	0.280	0.332	0.344	0.320
1p_tfidf	0.260	0.280	0.300	0.348	0.328

Quando considerados somente os documentos altamente relevantes, os resultados do desempenho da recuperação apresentam precisão média e P10 iguais a 0.249 e 0.252,

respectivamente. As tabelas 3 e 4 mostram essas precisões, quando a recuperação é feita com extratos. Como se vê, os resultados de P10 com extratos se aproximam bastante de P10 com indexação completa e foram em dois casos superiores a esse desempenho. Com esses resultados, Sakai e Sparck-Jones concluem que os extratos são úteis para recuperar documentos altamente relevantes, considerando-se os primeiros documentos recuperados. Já quando considerados os documentos julgados relevantes pela TREC, os extratos foram pouco úteis.

Tabela 3 – Precisão média com extratos para indexação, somente documentos altamente relevantes

Extrato	95%	90%	AL	70%	50%
lead	0.147	0.159	0.186	0.215	0.215
tfidf	0.155	0.166	0.198	0.216	0.236
1p_tfidf	0.144	0.162	0.192	0.219	0.233

Tabela 4 – P10 com extratos para indexação, somente documentos altamente relevantes

Extrato	95%	90%	AL	70%	50%
lead	0.240	0.252	0.260	0.248	0.248
tfidf	0.228	0.236	0.236	0.252	0.224
1p_tfidf	0.220	0.236	0.224	0.256	0.236

Brandow et al. verificaram a utilidade de dois sumarizadores automáticos: um que extrai sentenças baseando-se na medida tf-idf, e outro que as extrai de forma igual ao método *lead* descrito anteriormente. Os tamanhos dos extratos foram de 60, 150 e 250 palavras, independentemente do tamanho do documento.

Seus experimentos utilizaram um sistema de RI baseado no modelo booleano (detalhado no Capítulo 2) para 12 consultas a uma coleção com cerca de 20.000 documentos. Esse número é considerado pequeno quando comparado com as coleções atuais, que possuem mais de 50 mil documentos. Como resultado, a recuperação com indexação completa obteve

0.37 de precisão e 1.00 de revocação. A Tabela 5 mostra o desempenho das recuperações que usam os dois tipos de extratos.

Tabela 5 – Precisão e Revocação com extratos para indexação

Tamanho / Extrato	**60**		**150**		**250**	
	Precisão	Revocação	Precisão	Revocação	Precisão	Revocação
tfidf	0.45	0.42	0.46	0.59	0.46	0.67
lead	0.50	0.46	0.44	0.57	0.44	0.71

Como pode ser visto, todos os índices de precisão, independentemente do tamanho dos extratos, foram maiores que a precisão quando se efetuou a indexação completa.

Diferentemente da precisão, a revocação apresentou muitas variações para diferentes tamanhos de extratos. A revocação aumenta conforme o tamanho do extrato aumenta, porém, todas as taxas de revocação para a indexação por extratos foram menores que a revocação na indexação completa.

De um modo geral, os extratos aumentam a taxa de precisão de recuperação, comparada com a recuperação que utiliza os documentos completos, porém, há perda de revocação. De certa forma, essa perda com revocação é esperada pelo fato de o modelo booleano basear-se na rígida regra da presença de todos os termos da consulta. Como o extrato é geralmente menor que o documento, a chance dos termos não estarem presentes no extrato é maior; portanto, maior a chance do respectivo documento não ser recuperado. Para evitar esse declínio, Brandow et al. sugerem um sistema baseado em modelo diferente do booleano.

Seguindo a metodologia dos trabalhos descritos, neste trabalho também se usaram extratos na indexação, particularmente produzidos pelo GistSumm (PARDO, 2005; PARDO et al., 2003; PARDO, 2002), como será descrito no Capítulo 5.

Extratos foram também explorados na Realimentação de Pseudo-Relevantes, tópico abordado a seguir.

4 Realimentação de Relevantes

Muitas vezes usuários de sistemas de RI apresentam consultas que não são boas para recuperar documentos que podem atender às suas necessidades de informação. As razões podem ser a falta de termos na consulta ou, muitas vezes, o fato de o vocabulário utilizado pelos autores dos documentos ser diferente do utilizado nas consultas. Surge, então, a necessidade de melhorar essas consultas de alguma forma para que o usuário fique satisfeito com o resultado. Uma maneira de fazer isso é usando a Realimentação de Relevantes, ou RR (ROCCHIO, 1971).

A essência da RR é a seguinte: depois de uma busca inicial por documentos da coleção, os considerados mais relevantes pelo sistema são recuperados e suas descrições, apresentadas ao usuário (por exemplo, títulos, sumários, etc.). O usuário, então, examina essas informações e identifica quais os documentos relevantes e os irrelevantes à sua consulta. Poder-se-ia dizer que o conjunto de documentos irrelevantes é o complemento dos documentos relevantes, porém há casos em que o usuário não consegue ou não quer julgar a relevância de todos os documentos apresentados. Razão pela qual é importante o usuário notificar ao sistema para que o documento não seja usado de forma indevida na realimentação.

O sistema considera os documentos julgados pelo usuário e faz um ajuste na consulta original, promovendo alguns de seus termos e/ou adicionando outros de acordo com a resposta do usuário. Termos não relacionados a essa resposta podem também ser penalizados. A promoção e penalização referem-se ao incremento e decremento dos pesos dos termos da consulta original; esses dois processos são chamados de **reponderação da consulta**, enquanto a adição refere-se à inserção de novos termos que não estão presentes na consulta original,

processo denominado **expansão de consulta**. A realização de um ou ambos os processos é identificada na RI como **reformulação da consulta**.

Na reponderação, a idéia básica é que o julgamento do usuário deve tornar mais expressiva a consulta: os termos promovidos realçam a importância e os penalizados não devem ser preponderantes na recuperação dos documentos de uma coleção. A expansão de consulta serve para resolver as diferenças entre o vocabulário utilizado pelos autores nos documentos e o utilizado pelos usuários em suas consultas. Ela também serve para aprimorar consultas mal formuladas pelos usuários, quando eles não expressam adequadamente suas necessidades.

Após a reformulação da consulta original, um novo processo de busca pode ser repetido até que o usuário esteja satisfeito com a resposta ou até que nenhuma mudança adicional nos documentos recuperados seja observada (ELLIOTT; CASHMAN, 1973).

A reformulação de uma consulta pode ser expressa pela seguinte função (ROCCHIO, 1971):

$$Q_r = f(Q_0, R, S) \tag{13}$$

sendo Q_r a consulta reformulada, Q_0 a consulta inicial, R o conjunto de documentos julgados relevantes e S o conjunto dos julgados irrelevantes. No modelo vetorial, Q_r e Q_0 são vetores e R e S são conjuntos de vetores representando os documentos relevantes e irrelevantes, respectivamente. Esses vetores contêm os pesos dos termos na consulta ou no documento. Para $R = \{R_1, R_2, ..., R_{n1}\}$ e $S = \{S_1, S_2, ..., S_{n2}\}$a função de reformulação é descrita pela Eq. 14, em que a dimensão de todos os vetores é a mesma, pois cada vetor tem os pesos de todos os termos da coleção de documentos indexada.

$$Q_r = Q_0 + \frac{1}{n_1}\sum_{i=1}^{n1} R_i - \frac{1}{n_2}\sum_{i=1}^{n2} S_i \tag{14}$$

Quando a consulta ou documento não tiver um termo, o peso desse termo será nulo. Se a consulta introduz um novo termo que não está presente na coleção, esse termo é descartado, pois sua utilidade na recuperação é nula.

A Eq. 14 pode ser modificada para distinguir a influência dos termos da consulta original e dos documentos relevantes e irrelevantes, resultando na Eq. 15. Dessa forma, pode-se dar mais importância aos termos iniciais da consulta.

$$Q_r = \alpha Q_0 + \frac{\beta}{n_1} \sum_{i=1}^{n1} R_i - \frac{\gamma}{n_2} \sum_{i=1}^{n2} S_i \qquad (15)$$

Usando a Eq. 14 para reformular 17 consultas, Rocchio conduz um experimento que indica que a reformulação de consulta permite exibir documentos mais bem classificados que os documentos recuperados com a consulta original.

Já a expansão de consultas pode ser interativa ou automática. A expansão interativa exige que o usuário selecione os termos a serem acrescentados à sua consulta, enquanto a automática é feita totalmente pelo sistema. Um argumento a favor desta é que o sistema tem acesso a informações estatísticas sobre a utilidade relativa dos termos da expansão. O principal argumento a favor da expansão interativa é que ela dá mais controle ao usuário (RUTHVEN, 2003).

Além dos documentos julgados relevantes pelos usuários, há outras fontes para determinar a importância de um termo para uma consulta, como por exemplo: tesauros, relações de coocorrência entre palavras e documentos pseudo-relevantes a uma consulta. Documentos pseudo-relevantes são considerados como relevantes à consulta do usuário, porém não foram julgados pelo mesmo.

Mandala et al. (1999) propuseram o uso de três tipos de tesauros para expandirem suas consultas: tesauros construídos por humanos, nesse caso, a WordNet (MILLER, 1995);

tesauros construídos com base em coocorrências de palavras[14]; e construídos com base em palavras que coocorrem e que possuem específicas relações lingüísticas entre si (por exemplo, sujeito-verbo). Seus experimentos com a TREC 7 (VOORHEES; HARMAN, 1998) mostram que a expansão de consulta baseada nos três tesauros aumenta a precisão média da recuperação, quando comparada com a expansão que usa um só tipo de tesauro.

Xu e Croft (1996) fizeram expansão de consulta com palavras que coocorriam nos documentos das coleções utilizadas, fazendo o que eles chamaram de análise global. Seus experimentos com as coleções TREC 3 (HARMAN, 1995), TREC 4 (HARMAN, 1996) e WEST mostram que esse tipo de estratégia traz poucos ganhos para a efetividade, comparada à expansão que utiliza documentos pseudo-relevantes. A RR que considera documentos pseudo-relevantes, chamada de Realimentação de Pseudo-relevantes, será tratada na seção a seguir.

4.1 Realimentação de Pseudo-relevantes

A Realimentação de Pseudo-relevantes (RPR) é um tipo de RR que faz uso de documentos pseudo-relevantes, isto é, aqueles que estão no topo da lista de documentos recuperados automaticamente com a consulta original, ou seja, que possuem maior grau de similaridade com a consulta.

A principal diferença entre a RR original e a RPR é que, na última, a relevância dos documentos é determinada automaticamente, enquanto na primeira o usuário faz o julgamento. Segundo Lam-Adesina e Jones (2001), a RPR, em média, provê melhoras na performance da recuperação, mas os ganhos são menores que os observados na RR com julgamento humano. Essa diferença é compreensível, pois métodos automáticos são meras

[14] O uso de coocorrências, nesse trabalho, é baseado na suposição de que pares de palavras que ocorrem freqüentemente no mesmo documento sejam relacionadas ao mesmo assunto.

aproximações de casos reais de julgamentos humanos. Apesar do desempenho inferior, a RPR é recomendável em casos nos quais não se pode ter o julgamento dos usuários.

Na RPR, a relação entre a consulta original (Q_0), os documentos pseudo-relevantes (R) e a nova consulta (Q_r), geralmente, é dada por:

$$Q_r = f(Q_0, R) \tag{16}$$

Há abordagens que empregam a RPR considerando também documentos irrelevantes, que, na verdade, são pseudo-irrelevantes. Nesse caso, os documentos pseudo-irrelevantes são os que possuem um grau de similaridade muito baixo com a consulta inicial.

Na RPR há questionamentos importantes, tais como: quantos documentos devem ser considerados pseudo-relevantes, quantos e quais termos devem ser acrescentados e a maneira como a consulta deve ser reformulada.

O número ótimo de documentos pseudo-relevantes pode variar de acordo com o desempenho do sistema, com a coleção indexada, com as necessidades apresentadas pelos usuários e com a qualidade das consultas iniciais. Essas características podem influenciar a recuperação, podendo fazer com que documentos considerados pseudo-relevantes sejam de fato, irrelevantes. Dessa forma, os termos selecionados podem fazer com que a recuperação que usa a consulta reformulada contenha um número maior de documentos irrelevantes que a recuperação inicial, diminuindo, portanto a precisão do sistema. É necessário considerar um número de documentos pseudo-relevantes pequeno quando o sistema tende a recuperar poucos documentos relevantes na busca inicial.

O padrão da consulta apresentada ao sistema também pode influenciar muito a escolha do número de pseudo-relevantes. Em casos em que a consulta é muito genérica, um número baixo de documentos pseudo-relevantes poderá limitar muito a consulta do usuário, ou seja, a consulta irá tornar-se muito específica, contrariando a intenção do usuário.

Em experimentos com a TREC, Montgomery et al. (2004) apresentaram um experimento com nove sistemas de RI. Cada sistema utilizou 20 termos na reformulação de consultas e o número de documentos pseudo-relevantes variou de 1 a 100. Como resultado, o número ideal[15] variou muito: enquanto alguns sistemas tinham a performance reduzida rapidamente de acordo com o acréscimo do número de documentos considerados, outros pareciam insensíveis à adição. No entanto, em geral, os sistemas apresentaram uma boa performance quando consideraram um número limitante de documentos pseudo-relevantes. Após esse limite, a reformulação de consulta passou a ter um efeito negativo. Em média, o limite de 15 documentos apresentou ganhos na performance.

Além de escolher o número de documentos a serem considerados como relevantes, há estratégias para selecionar os termos a serem usados na reformulação da consulta, descritas a seguir.

4.1.1 Métodos de seleção automática de termos

Na RPR, a reformulação da consulta se dá pela extração dos termos contidos nos documentos pseudo-relevantes, mas não necessariamente todos são utilizados, já que essa utilização (como acontece quando se utiliza a fórmula original de Rocchio) pode ser um grande erro porque alguns termos podem prejudicar a efetividade da recuperação efetuada com a nova consulta. Esse prejuízo tende a ocorrer principalmente quando os documentos utilizados na reformulação possuem vários tópicos muitos dos quais não são relevantes à consulta inicial. Uma estratégia é selecionar apenas alguns dos termos dos documentos pseudo-relevantes. Harman (1992) em um experimento com uma pequena coleção chamada

[15] Número ideal é o número que faz com que a recuperação do sistema seja mais efetiva. Nesse experimento, após excedido um certo número (diferente para cada sistema) a performance dos sistemas usando RPR diminuía.

Cranfield 1400[16], mostra que os resultados com 20 termos adicionados à consulta são melhores que quando todos os termos dos documentos são acrescentados.

Para selecionar os termos que poderão promover uma melhora na recuperação, várias medidas já foram utilizadas em pesquisas anteriores. Uma medida bem simples é utilizar o somatório do peso do termo nos documentos relevantes, pois ele pode indicar os termos mais importantes para a reformulação. Outras medidas, tais como a χ^2 (qui-quadrado) e a *Robertson selection value* , ou rsv (ROBERSTON, 1990), são baseadas na distribuição dos termos nos documentos relevantes e nos documentos da coleção indexada. O valor χ^2 para o termo t é determinado pela seguinte equação:

$$\chi^2(t) = \frac{\left(p_r(t) - p_c(t)\right)^2}{p_c(t)} \qquad (17)$$

em que $p_r(t)$ é a probabilidade de ocorrência do termo t nos documentos relevantes (ou pseudo-relevantes) e $p_c(t)$ é a probabilidade de ocorrência do termo t na coleção toda. O valor χ^2 é proporcional à diferença entre a probabilidade de ocorrência do termo t nos documentos relevantes e a probabilidade de ocorrência do termo t na coleção. Dessa forma, se t tem alta probabilidade de ocorrência na coleção e também nos documentos relevantes, o valor χ^2 será baixo; caso t tenha baixa probabilidade de ocorrência na coleção, mas alta nos documentos relevantes, o valor χ^2 será alto. Então, χ^2 privilegia termos com alta concentração nos documentos relevantes.

Carpineto e Romano (1999) estimam essas duas probabilidades da seguinte forma:

$$p_r(t) = \frac{fr(t)}{n(tr)} \qquad (18)$$

$$p_c(t) = \frac{fc(t)}{n(tc)} \qquad (19)$$

[16] A Cranfield 1400 é uma coleção de testes que contém 1400 sumários manuais relacionados à aeronáutica e 225 consultas juntamente com julgamentos das relevâncias dos sumários para as consultas.

em que fr(t) é a freqüência do termo nos documentos relevantes, n(tr) é o número de termos nos documentos relevantes, fc(t) é a freqüência do termo na coleção, enquanto n(tc) é o número de termos na coleção. Dessa forma, $p_r(t)$ é a proporção de termos t no conjunto de documentos relevantes e $p_c(t)$ é a proporção de termos t na coleção de documentos.

A rsv é expressa por:

$$\mathrm{rsv}(t) = \mathrm{r}(t) * \mathrm{rw}(t) \tag{20}$$

na qual r(t) é o número de documentos relevantes contendo o termo t e rw(t) é o índice de relevância do termo t, que pode ser determinado pela seguinte fórmula (ROBERTSON; SPARCK-JONES, 1976):

$$\mathrm{rw}(t) = \log\left(\frac{(\mathrm{r}(t)+0.5)*(\mathrm{N}-\mathrm{n}(t)-\mathrm{R}+\mathrm{r}(t)+0.5)}{(\mathrm{n}(t)-\mathrm{r}(t)+0.5)*(\mathrm{R}-\mathrm{r}(t)+0.5)}\right) \tag{21}$$

em que N é o número total de documentos indexados, R é o número total de documentos relevantes, n(t) é o número de documentos contendo o termo t e r(t), o número de documentos relevantes contendo o termo t. O índice de relevância é calculado pela diferença entre a probabilidade do termo t ocorrer em um documento relevante ($p_{rel}(t)$, Eq. 22) e a probabilidade do termo t pertencer a um documento não-relevante ($p_{nrel}(t)$, Eq. 23). Devido a problemas introduzidos por baixos valores de R e r(t), que ocorrem freqüentemente na prática, o fator de ajustamento 0.5 é introduzido em cada membro das probabilidades $p_{rel}(t)$ e $p_{nrel}(t)$. Valores críticos são, por exemplo, R=1 e r(t)=1 ou R=1 e r(t)=0, pois, no primeiro caso ocorrerá uma divisão por 0 e, no segundo caso, ocorrerá um valor inválido para a função de logaritmo. Dessa forma, a rsv irá privilegiar termos que ocorrem na maioria dos documentos relevantes e que têm maior probabilidade de ocorrer nesses documentos do que no restante da coleção.

$$p_{rel}(t) = \log\left(\frac{r(t)+0.5}{R-r(t)+0.5}\right) \quad (22)$$

$$p_{nrel}(t) = \log\left(\frac{n(t)-r(t)+0.5}{N-R-n(t)+r(t)+0.5}\right) \quad (23)$$

Após o cálculo dos índices de relevância de termos dos documentos por uma das medidas descritas, basta definir o número de termos a serem usados na reformulação de consulta e selecionar aqueles com maior pontuação. Em geral, os experimentos com as coleções CLEF e TREC utilizam 5, 10 ou 20 termos na reformulação da consulta, como os de Llopis et al. (2004), Sakai e Sparck-Jones (2001) ou Adrafe et al. (2004).

Carpineto e Romano (1999) apresentam uma estratégia para reformular a consulta considerando os valores atribuídos pelas medidas de seleção automática de termos, tais como a rsv. A estratégia é expressa pela seguinte fórmula de ponderação:

$$w(t)_{qr} = \alpha * w(t)_{q0} + \beta * mvalor(t) \quad (24)$$

para $w(t)_{qr}$ como o peso do termo na consulta reformulada, $w(t)_{q0}$ como o peso do termo na consulta inicial, mvalor(t) como o valor atribuído ao termo t pela medida de seleção de termos (se o termo não foi selecionado o valor será nulo) e α e β determinados conforme a estimativa da qualidade da recuperação inicial. Se a recuperação inicial possuir poucos documentos relevantes, a diferença entre α e β será alta. Por exemplo, para a TREC 9, que, segundo Carpineto e Romano, é uma coleção com recuperação inicial pobre, α e β são iguais a 1 e 0.2, respectivamente. Já para a TREC 8, Carpineto e Romano determinaram valores iguais a 1 para α e β, pois a recuperação inicial é boa. O valor de β menor que α irá fazer com que os novos termos influenciem menos na recuperação final**.** Nos sistemas desenvolvidos neste mestrado, foram adotados os mesmos valores de α e β que Carpineto e Romano utilizaram para processar a TREC 9, pois a recuperação inicial também apresentou poucos documentos relevantes.

Nesta seção mostrou-se como os termos de documentos pseudo-relevantes podem ser selecionados e como os selecionados podem ser ponderados para que a consulta original seja reformulada e, assim, a RPR seja efetuada. Na próxima seção, mostrar-se-ão pesquisas que consideram o uso de uma importante fonte para seleção de termos de um documento pseudo-relevante: o sumário.

4.2 Emprego de Sumarização Automática na Realimentação de Pseudo-relevantes

A utilização de sumários é justificada na seleção automática de termos da RPR, pois um sumário é composto pela seleção das principais partes de um documento. Estas, por sua vez, podem conter termos potencialmente relevantes para a reformulação de consultas. Os mesmos métodos de seleção de termos apresentados na seção anterior podem ser usados quando os sumários são as fontes de termos para a reformulação da consulta (Figura 7).

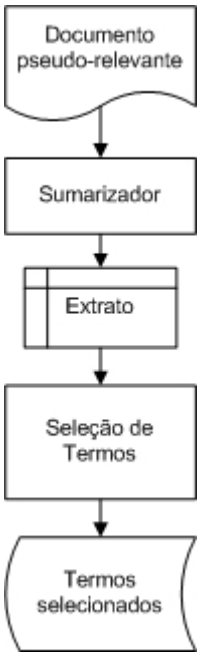


Figura 7 – Uso de sumários na seleção de termos da RPR

Como mostra a figura, a seleção de termos passa a ter duas etapas. Primeiramente, são gerados os extratos e só depois ocorre a seleção dos termos que serão usados na reformulação

da consulta. A etapa de sumarização pode não ocorrer em casos em que os documentos pseudo-relevantes possuem sumários manuais, pois esses mesmos podem ser usados como fontes de seleção de termos.

A seguir, os trabalhos de Lam-Adesina e Jones (2001) e Sakai e Sparck-Jones (2001), que têm como foco a utilização de sumários na RPR, são apresentados.

4.2.1 Utilização de extratos por Lam-Adesina e Jones

Lam-Adesina e Jones (2001) notaram que termos não-relevantes a uma necessidade de informação são, erroneamente, utilizados na reformulação das consultas, pelo fato de pertencerem a documentos pseudo-relevantes. Para evidenciar esse fato, apresentam como exemplo uma consulta cuja necessidade de informação é saber qual é o impacto econômico da reciclagem de pneus. A maioria dos termos selecionados e utilizados no refinamento da consulta foi irrelevante, ou seja, não se referiu ao tópico da consulta. Por exemplo, havia termos referentes a outros materiais recicláveis, tais como plástico e vidro. Apesar disso, eles foram automaticamente adicionados à consulta, prejudicando a recuperação.

Para resolver esse problema de adição de termos irrelevantes, Lam-Adesina e Jones propõem o uso de extratos guiados pela consulta original, ou seja, específicos, como fonte de seleção de termos. Esses extratos são formados por sentenças extraídas dos documentos pseudo-relevantes. Essas sentenças são aquelas que obtêm as maiores pontuações, de acordo com a combinação dos seguintes métodos de pontuação: freqüência dos termos da sentença no documento, presença de termos do título do documento na sentença, localização da sentença no documento (ressaltando a importância das duas primeiras sentenças em artigos jornalísticos) e a presença de termos da consulta na sentença.

Com a finalidade de comparação, abordagens que consideram termos do documento e do respectivo extrato genérico também foram consideradas. Para a geração dos extratos

genéricos também foram consideradas as sentenças com maior pontuação, obtidas pelos mesmos métodos usados para criar extratos específicos, mas excluindo-se o que considera a presença de termos da consulta (já que são extratos genéricos).

A taxa de compressão dos extratos também é importante: ela deve ser suficiente para conter um número satisfatório de termos relevantes para a expansão, porém, não tão baixa que faça com que o extrato possua um número considerável de termos irrelevantes. Os extratos utilizados nos experimentos de Lam-Adesina e Jones foram gerados com uma taxa de compressão de 85%. Na fase de pré-processamento, eles fazem a remoção de *stopwords* e o *stemming*. O sistema faz a RPR com extratos dos cinco primeiros documentos recuperados, como fonte de termos candidatos à reformulação de consulta.

Dos extratos são selecionados 20 termos com a medida rsv (Eq. 20, p. 50), que foi considerada a melhor medida disponível. Deve-se destacar que, para o cálculo da rsv, foram consideradas as ocorrências dos termos nos documentos. Os extratos foram utilizados somente para filtrar os termos relevantes à reformulação.

A coleção TREC 8 (VOORHEES; HARMAN, 1999) foi utilizada para testar a metodologia, com 538.151 documentos das seguintes fontes: *Federal Register*, *Finacial Times*, *Foreign Broadcast Information Service e LA Times*. Para a formação automática das consultas, foi considerado somente o campo *title* do conjunto dos tópicos da TREC 8.

Os experimentos com esta coleção mostraram que a reformulação com seleção por extratos, específicos ou genéricos, apresenta uma melhora na precisão dos resultados acima de 15% relativa ao *baseline* (o sistema sem RPR). Bem como uma melhora na precisão dos resultados acima de 11%, se comparada ao desempenho da RPR, quando todos os termos dos documentos pseudo-relevantes foram considerados candidatos à seleção.

A seleção usando extratos específicos também mostra melhores desempenhos que os extratos genéricos, porém, as diferenças são pequenas. Isso mostra que o extrato genérico

poderia ser utilizado quando o específico não pode ser empregado por restrições do ambiente de aplicação ou por questões de eficiência. Os extratos genéricos podem, inclusive, ser gerados em uma fase anterior a interação do usuário com o sistema.

A reformulação de consulta que utiliza extratos específicos mostra ganho na efetividade comparada ao desempenho do *baseline*, mesmo quando os documentos pseudo-relevantes são de fato irrelevantes. Já considerando os documentos correspondentes, em vez de extratos, a performance foi degradada. Isso mostra que os extratos podem reduzir o impacto dos termos irrelevantes contidos em documentos irrelevantes.

Essas considerações de Lam-Adesina e Jones foram baseadas na precisão média e precisão para os primeiros 10 e 30 documentos recuperados, P10 e P30 (detalhadas no Capítulo 2). Eles concluíram que extratos específicos são boas fontes para termos candidatos à reformulação de consulta.

4.2.2 Uso de sumários e extratos por Sakai e Sparck-Jones

Sakai e Sparck-Jones (2001) investigaram se sumários genéricos podem melhorar a efetividade na RI, empregando metodologia bastante similar à de Lam-Adesina e Jones (2001). As taxas de compressão, as formas de geração de extratos e a coleção utilizada na avaliação são aquelas apresentadas no Capítulo 3, quando relatados seus experimentos com sumários na indexação.

As abordagens usadas para verificar o desempenho dos sumários na RPR são as seguintes:

- Sumário-Sumário: sumários utilizados como índice tanto na busca inicial, quanto na final e na RPR.

- Texto-Texto: documentos utilizados como índice na busca inicial e final e na RPR.
- Sumário-Texto: sumários utilizados como índice na busca inicial e na RPR, e documentos como índice na busca final.
- Sumário como filtro: documentos utilizados como índice na busca inicial e final e na RPR, mas os sumários como filtro de seleção de termos. Essa abordagem é a semelhante àquela empregada por Lam-Adesina e Jones.

Nessas abordagens, utilizar o documento ou o sumário na RPR indica que os valores da medida de seleção são computados com base nos documentos ou sumários, respectivamente. Utilizar o sumário como filtro de seleção de termos indica que somente os termos presentes nos sumários são candidatos à expansão. Assim, a abordagem 'Sumário como filtro' utiliza os termos dos sumários como candidatos, porém computa a medida de seleção baseando-se nas ocorrências dos termos nos documentos.

Na RPR, os documentos considerados pseudo-relevantes são os cinco primeiros da lista de documentos recuperados (ou seja, aqueles com maiores graus de similaridade com a consulta). Um conjunto de 30 termos foi adicionado à consulta inicial. A medida utilizada para a determinação de relevância dos termos foi a rsv (Eq. 20, p. 50).

Nos experimentos com a abordagem Sumário-Sumário, constatou-se que a precisão aumenta com o aumento do tamanho dos extratos (taxas de compressão de 95%, 90%, 70% e 50%), independentemente do tipo de extrato (*lead,* tfidf ou 1p_tfidf). Ao considerar a precisão para os primeiros 10 documentos recuperados (P10), as taxas de 90 e 70% apresentaram os melhores valores para quase todos os tipos de extrato.

Considerando as duas medidas (P10 e precisão média), os resultados da abordagem Texto-Texto são superiores à abordagem que utiliza o documento como índice, sem a etapa de RPR. No entanto, a diferença não é muito significativa. A abordagem Sumário-Sumário, na

maioria das vezes, supera a recuperação utilizando o sumário como índice, sem RPR. Isso mostra que a RPR traz melhores resultados na recuperação, na maioria dos casos. A abordagem Sumário-Sumário só supera Texto-Texto na medida P10, quando se consideram os documentos altamente relevantes e quando os extratos têm taxas de compressão de 70% ou quando os sumários são manuais.

A abordagem Sumário-Texto apresenta, em média, valores de P10 superiores à recuperação que utiliza a abordagem Texto-Texto, independentemente da taxa de compressão e do tipo de extração, considerando todos os documentos relevantes e também os altamente relevantes. A abordagem 'Sumário como filtro', utilizando extratos com taxa de compressão de 70% também supera a abordagem Texto-Texto, independentemente do tipo de extração. Essa abordagem, quando considera extratos com taxa de compressão de 50%, obtém a mesma precisão que a abordagem Texto-Texto, para todos os tipos de extratos e documentos relevantes e altamente relevantes. Para outras taxas de compressão o desempenho foi inferior.

Para a precisão média, a abordagem Sumário-Texto obteve desempenho inferior à Texto-Texto, na maioria dos casos. Em média, a abordagem 'Sumário como Filtro' é superior à abordagem Sumário-Texto, quando se considera a precisão média, e esta é superior à 'Sumário como filtro', quando se considera P10. Isso indica que, se o objetivo for obter a melhor precisão média, a abordagem mais indicada é 'Sumário como filtro'. Mas, se for obter a melhor P10, a abordagem mais indicada é Sumário-Texto.

Os resultados ainda mostram que a abordagem Sumário-Texto é significativamente mais efetiva que a Sumário-Sumário. Sakai e Sparck-Jones concluem, assim, que a sua hipótese de que sumários são úteis na RPR é comprovada, principalmente quando considerada a abordagem Sumário-Texto. Essa abordagem mostra, assim, ganhos em efetividade, considerando a medida P10 para documentos relevantes e altamente relevantes, independentemente do tipo de extrato.

No próximo capítulo serão apresentados os sistemas desenvolvidos neste trabalho para explorar o emprego de extratos tanto na indexação quanto na RPR.

5 Recuperação de Informação com Auxílio de Extratos Automáticos

Neste trabalho, foram construídos vários sistemas de RI para recuperar documentos em português, com o intuito de verificar a utilização de extratos tanto na etapa de indexação quanto na etapa de RPR. A diferença com relação aos trabalhos citados anteriormente é o sumarizador utilizado para a geração dos extratos, o GistSumm (PARDO, 2005; PARDO et al., 2003; PARDO, 2002), que é um sumarizador extrativo já disponível no Núcleo Interinstitucional de Lingüística Computacional (NILC)[17]. Duas hipóteses são investigadas:

i) os índices construídos a partir de extratos do GistSumm podem ser tão efetivos quanto os produzidos a partir de documentos.

ii) os extratos do GistSumm podem aumentar a efetividade da recuperação quando utilizados na RPR.

A escolha do GistSumm é baseada na utilidade dos seus extratos para indicar a relevância de um documento para o leitor, constatada por avaliação na Document Understanding Conference 2003 (DUC-2003). Segundo Pardo et al. (2002), o GistSumm tenta simular o comportamento humano na tarefa de sumarização observado quando uma pessoa tenta identificar os tópicos principais do texto e reestruturá-lo, usando informações complementares quando necessárias para produzir um texto coeso. O GistSumm é baseado em duas premissas. A primeira é a de que todo texto contem uma idéia principal. A segunda é a de que deve ser possível identificar uma única sentença que melhor expresse essa idéia, a sentença *gist*. O modo como o GistSumm seleciona a sentença *gist* e as demais que irão compor o extrato será detalhado posteriormente.

[17] http://www.nilc.icmc.usp.br; último acesso em 13/06/2006.

As seções a seguir descrevem os sistemas construídos neste trabalho. A primeira detalha um sistema de RI básico, que não usa extratos em nenhuma de suas etapas de recuperação, o RDoc. A segunda descreve dois sistemas que utilizam extratos genéricos na indexação, o RExt e o RDocExt. A terceira apresenta sistemas que utilizam extratos na RPR, cada um deles explorando um tipo diferente de extratos, a saber: extratos genéricos mono-documento, específicos mono-documento e específicos multi-documentos.

5.1 RDoc: Recuperador padrão

O sistema RDoc foi criado para ser o *baseline* para a avaliação dos outros sistemas. Como *baseline*, ele não utiliza extratos em nenhuma das etapas do processo de recuperação. Ao contrário, ele é um exemplar de um recuperador padrão, cuja indexação é feita diretamente com os documentos da coleção em foco. Além disso, a exibição de resultados da busca é direta, sem qualquer reformulação da consulta. A arquitetura do RDoc é mostrada na Figura 8.

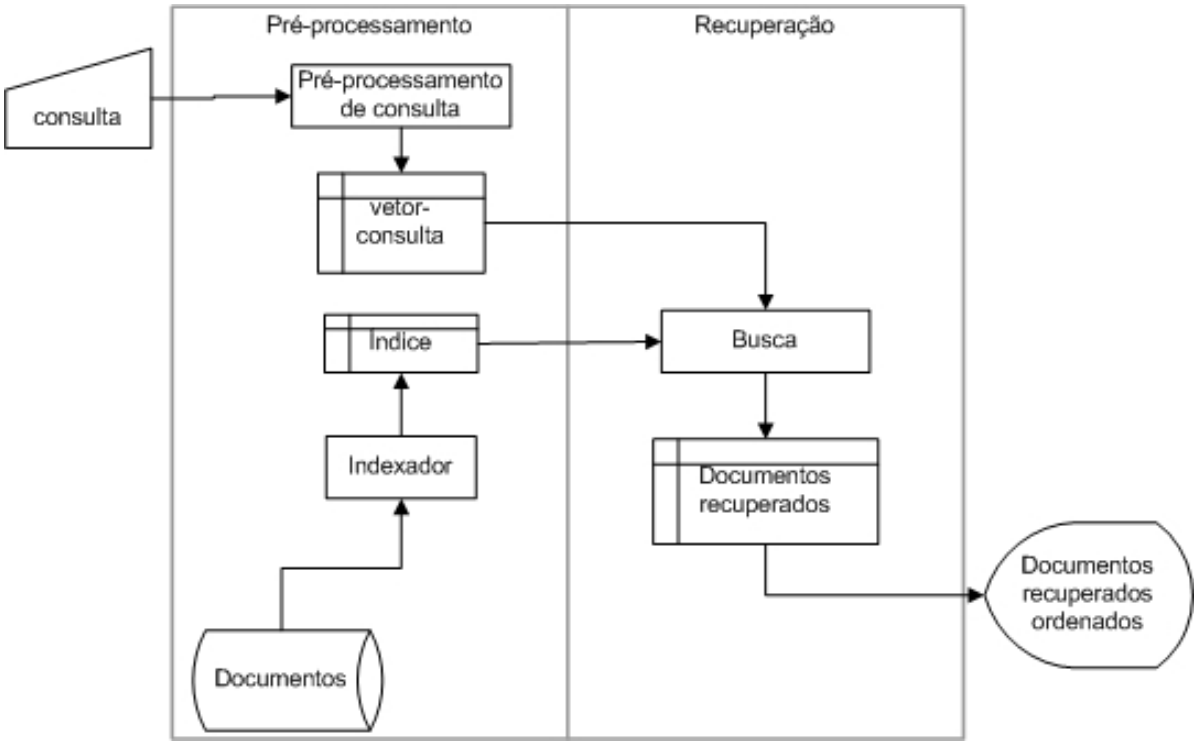


Figura 8 – Arquitetura do RDoc

Primeiramente, o arquivo de índices invertido é criado e armazenado. O indexador realiza as tarefas padrão de pré-processamento dos documentos, já descritas no Capítulo 3: tokenização, eliminação de etiquetas SGML, eliminação de *stopwords*, *stemming* e construção do arquivo de índices invertido. As etiquetas SGML são eliminadas porque não servem como termos indexadores de um documento. Os *tokens* que não são sinais de pontuação ou números passam pelo processo de *stemming*. O *stemmer* de Orengo e Huyck (2001) foi usado em todos os sistemas aqui apresentados, devido aos seus ganhos apresentados no Capítulo 3. Após esse pré-processamento, o arquivo de índices invertido é criado.

Uma vez criado o arquivo de índices invertido, o sistema pode efetuar uma busca a partir de uma consulta apresentada pelo usuário. A consulta, em língua natural, é pré-processada, resultando no vetor-consulta, que contém os radicais das palavras da consulta e seus respectivos pesos. Esse pré-processamento é semelhante ao processo de indexação. Primeiramente, a consulta em linguagem natural é tokenizada e as *stopwords* e sinais de pontuação são removidos. Após isso, cada *token* passa pelo processo de *stemming*. Os radicais

distintos resultantes são armazenados e a cada um é atribuído um peso. Seguindo Salton e Buckley (1988), o peso de um termo é igual ao produto da freqüência do termo na consulta pela medida idf do termo (Eq. 8, p. 35).

Na busca, os documentos com maiores graus de similaridade com o vetor-consulta são classificados em ordem decrescente de similaridade, formando o conjunto de documentos recuperados. A quantidade de documentos recuperados poderá variar dependendo da consulta, porém um limite é estabelecido previamente no sistema.

Neste trabalho, para calcular a similaridade entre o vetor-consulta e os documentos da coleção, foi usado o coeficiente de similaridade de Dice (Eq. 2, p. 8). Os pesos dos termos nos documentos foram determinados de acordo com as medidas tf-idf (Eq. 10, p. 36) e a que usa o logaritmo da freqüência (Eq. 12, p. 37). Para os demais sistemas apresentados neste capítulo o coeficiente de Dice também é utilizado, mas para ponderação dos termos somente a medida que usa o logaritmo da freqüência foi utilizada. A justificativa para não se utilizar a medida tf-idf na ponderação dos termos é apresentada no próximo capítulo.

Os outros sistemas propostos neste trabalho são variações dessa forma padrão de recuperação, como apresentado a seguir.

5.2 Utilizando extratos do GistSumm na indexação de documentos

Exploraram-se duas formas de utilização de extratos na indexação: somente extratos genéricos (sistema RExt) e extratos genéricos juntamente com os documentos (sistema RDocExt), constituindo uma indexação mista. Esses dois sistemas são apresentados nesta seção.

5.2.1 RExt: RI com indexação baseada em extratos

A principal motivação para a construção do RExt foi a de verificar se os extratos do GistSumm usados na indexação poderiam proporcionar uma efetividade semelhante à obtida pelo sistema que usa documentos na indexação, o RDoc.

A única variação do RExt (Figura 9) com relação ao RDoc é que o primeiro utiliza extratos genéricos na etapa de indexação, enquanto o segundo utiliza os documentos completos. Extratos genéricos são aqueles construídos com base nos possíveis tópicos principais de cada documento, sendo estes os que, supostamente, o autor mais enfatiza.

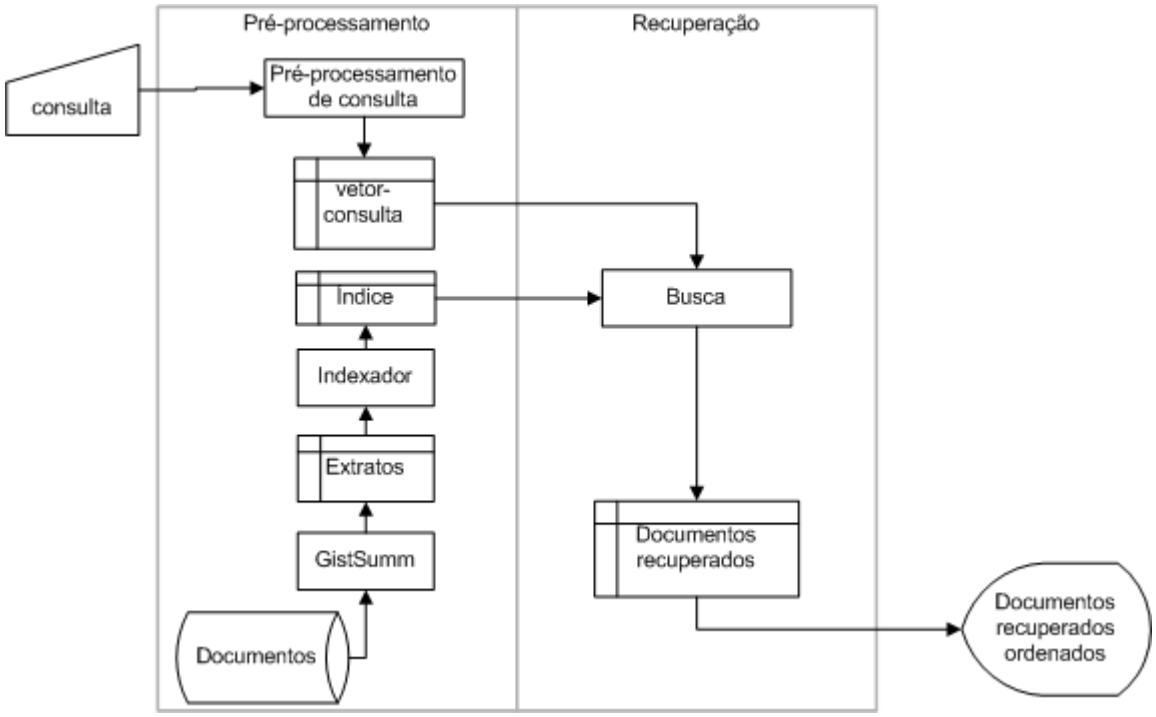


Figura 9 – Arquitetura do RExt

Para a indexação baseada em extratos, cada documento da coleção é primeiramente sumarizado pelo GistSumm. Esse processo gera uma coleção de extratos, que é então

indexada, formando o arquivo de índices invertido. Após a criação desse arquivo, o sistema está pronto para realizar buscas a partir de uma consulta apresentada por algum usuário.

Para gerar extratos genéricos, o GistSumm calcula a freqüência de todos os radicais das palavras contidas no documento e, então, determina uma pontuação para cada sentença baseada no somatório das freqüências de seus radicais no documento. A sentença que conseguir maior pontuação no documento é considerada a sentença *gist*. Outras sentenças são selecionadas para compor o extrato sob duas condições: (i) se alguns dos seus radicais ocorrerem na sentença *gist*; (ii) se a pontuação delas, obtida no processo de seleção da *gist*, for superior à média de pontuação das sentenças do documento. Independentemente da taxa de compressão, a sentença *gist* sempre será selecionada para compor o extrato.

Para avaliar a contribuição do RExt à RI, foram geradas duas coleções de extratos, com 60 e 80% de compressão em relação aos documentos da coleção original. Dessa forma, o RExt deu origem a duas versões: RExt60 e RExt80. A taxa de 60% foi escolhida devido ao fato de os resultados de Sakai e Sparck-Jones (2001) mostrarem que extratos com taxas de compressão próximas a 60% traziam os melhores resultados. A taxa de 80% foi adotada para verificar qual seria o comportamento quando a taxa de compressão fosse aumentada. Em muitos casos, essa taxa de compressão faz com que os extratos contenham somente a sentença *gist*.

5.2.2 RDocExt: RI com indexação mista

A motivação para a criação do RDocExt (Figura 10) é superar a precisão de um sistema de RI tradicional (RDoc) e, ao mesmo tempo, manter uma taxa de revocação aproximada.

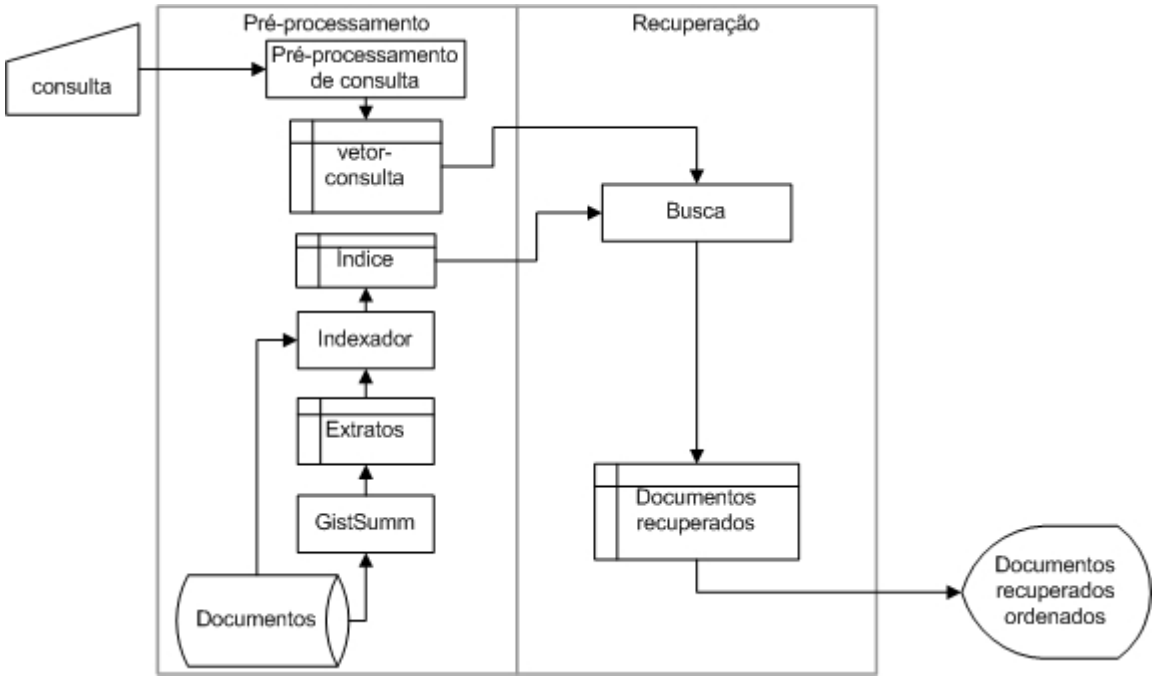


Figura 10 – Arquitetura do RDocExt

Na indexação mista, a ocorrência dos termos tanto no documento quanto nos extratos é considerada. Para o cálculo do peso do termo, é considerada a soma da freqüência do termo no documento com a freqüência do termo no respectivo extrato. O extrato é gerado com taxa de compressão de 80%, que foi escolhida para que seja possível uma comparação com o RExt. Já a taxa de compressão de 60% não foi considerada, pois, é incomum na SA, muito embora alguns trabalhos citados anteriormente tenham apresentados bons resultados com taxas próximas a 60%. Os processos de busca e processamento das consultas são similares aos do RDoc.

5.3 Utilizando extratos do GistSumm na Realimentação de Pseudo-relevantes

O principal motivo para o uso de extratos na RPR, em vez de documentos completos, é a de que há mais chance, no segundo caso, de introduzirem-se termos não relevantes à necessidade de informação (termos ruins) na consulta reformulada do que no primeiro. Para comprovar isso, o GistSumm se apresenta como um bom candidato, já que seu método de

construção de extratos seleciona somente as sentenças mais relacionadas à sentença *gist.* Em teoria, portanto, os termos ruins estarão incorporados às sentenças consideradas irrelevantes para o extrato.

Foram considerados extratos genéricos mono-documento e extratos específicos mono e multi-documentos. Também para fim de testes, foi criada uma abordagem que faz uso diretamente dos documentos na RPR. Os quatro sistemas de RI assim construídos chamam-se:

i) RFGenS – usa extratos mono-documento genéricos na RPR (***R**elevance **F**eedback using **Gen**eric **S**ingle-document Extracts*)

ii) RFQBS – usa extratos mono-documento específicos na RPR (***R**elevance **F**eedback using **Q**uery-**B**iased **S**ingle-document Extracts*);

iii) RFQBM – usa extratos multi-documentos específicos na RPR (***R**elevance **F**eedback using **Q**uery-**B**iased **M**ulti-documents Extracts*);

iv) RFFullDoc – usa os próprios documentos na RPR (***R**elevance **F**eedback using **Full Doc**uments*).

A arquitetura básica dos sistemas que usam extratos na RPR é apresentada na Figura 11. Como pode ser verificado, se for a primeira ocorrência da busca, deverá haver a RPR. Nesse caso, são gerados os extratos dos cinco documentos mais relevantes à consulta, já recuperados. Foi utilizada a taxa de compressão de 90%, pois se decidiu selecionar uma taxa um pouco maior que Lam-Adesina e Jones (2001), devido ao número de termos utilizados para a reformulação das consultas ser menor nos sistemas aqui desenvolvidos. Os extratos darão origem, então, à reformulação da consulta, a partir da qual a resposta final da busca será produzida.

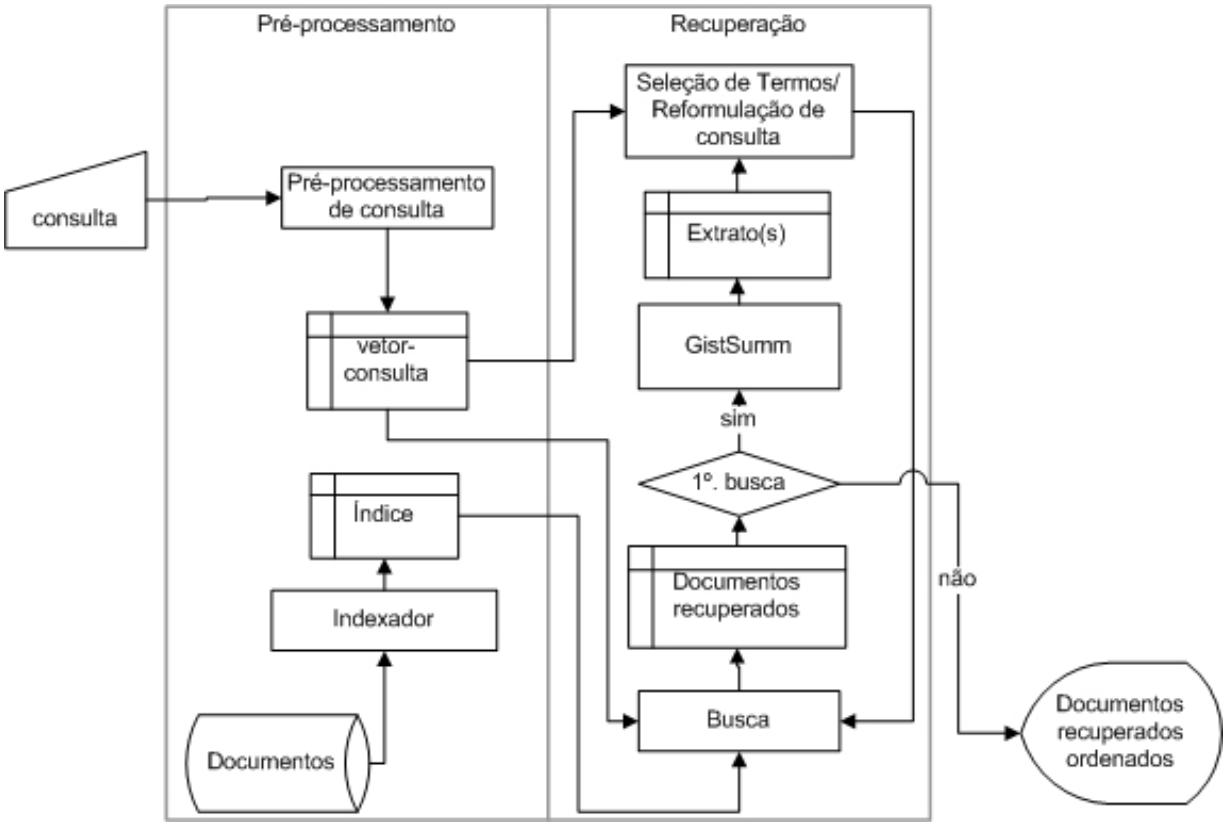


Figura 11 – Arquitetura geral dos sistemas que usam extratos na RPR

Na reformulação, os primeiros cinco documentos com maior grau de similaridade foram considerados pseudo-relevantes e seus extratos foram utilizados na reformulação da consulta. Nos casos em que a primeira busca retorna menos de cinco documentos, todos eles são considerados pseudo-relevantes. Essa quantidade foi escolhida em razão dos resultados obtidos por Montgomery et al. (2004), que mostraram que vários sistemas demonstravam ganho de performance contínua quando até os primeiros 15 documentos eram considerados pseudo-relevantes. Não foi considerado um número maior de documentos devido à queda de eficiência: quanto maior o número de documentos maior é o tempo de processamento.

Dez termos dos extratos são selecionados para a reformulação da consulta – aqueles com maiores valores de rsv (Eq. 20, p. 50). Essa escolha se deu porque Harman (1992) demonstrou que sistemas que utilizaram esse número na RPR conseguiram ganhos em efetividade. Nos experimentos de Lam-Adesina e Jones (2001), foi utilizado um número maior de termos na reformulação da consulta. Um número menor foi adotado aqui devido ao

fato de a recuperação inicial na coleção a ser utilizada nos experimentos retornar poucos documentos relevantes dentre os pseudo-relevantes. Caso os extratos tenham menos que dez termos, todos os termos são selecionados.

Após a escolha dos termos, os novos, ou seja, aqueles que não pertenciam à consulta inicial, são adicionados à consulta. A Eq. 24 (CARPINETO; ROMANO, 1999) é utilizada para ponderar esses termos novos e também para reponderar os termos da consulta inicial. A Eq. 24 é reproduzida abaixo:

$$w(t)_{qr} = \alpha * w(t)_{q0} + \beta * mvalor(t)$$

em que, nos sistemas, mvalor(t) é o valor rsv para o termo t. O $w(t)_{q0}$ é o peso do termo t na consulta inicial e $w(t)_{qr}$ é o peso do termo na consulta reformulada. Como Carpineto e Romano determinaram uma alta diferença entre α e β para a TREC 9, porque é uma coleção em que a recuperação inicial tende a retornar poucos documentos relevantes, o mesmo foi feito aqui, já que a recuperação inicial com a coleção CHAVE também é pobre. Os valores adotados foram 1 e 0.2 para α e β, respectivamente.

Uma vez reformulada, a consulta resultante dá origem ao processo real de busca na coleção de documentos, resultando nos documentos mais relevantes como resposta final ao usuário.

Esse processo é idêntico em todos os sistemas que usam extratos na RPR. A única variação entre eles, já enfatizada, é a forma como os extratos são gerados. Dessa forma, mudam-se as fontes de realimentação de pseudo-relevantes na consulta em cada um dos sistemas, descritos a seguir.

5.3.1 RFGenS: RPR com extratos genéricos mono-documento

O RFGenS utiliza os extratos genéricos produzidos pelo GistSumm na RPR. Como já dito na seção anterior, eles são construídos simplesmente visando aos tópicos principais do documento, guiados pela sentença *gist.* A motivação para usar esse tipo de extrato na RPR é que, muitas vezes, termos relacionados a informações periféricas do documento podem ser considerados relevantes e usados para reformular a consulta, produzindo uma consulta pobre ou até pior que a consulta original. Considerando-se os extratos com conteúdo relacionado aos tópicos principais dos documentos, haveria menor chance de se ter informação periféricas, portanto.

5.3.2 RFQBS: RPR com extratos específicos mono-documento

A RPR do RFQBS faz uso de extratos específicos para a realimentação da consulta de maneira muito semelhante à de Lam-Adesina e Jones (2001). Para produzi-los, o GistSumm tenta encontrar a sentença mais similar à consulta inicial, em vez de, simplesmente, calcular a sentença *gist* do documento. A sentença obtida será, assim, a sentença *gist* do documento com relação à consulta. A similaridade é calculada de acordo com o cosseno de similaridade (SALTON; McGILL, 1983). Outras sentenças serão incluídas no extrato respeitando-se sua similaridade com a consulta, a existência de termos comuns com a sentença *gist* e a taxa de compressão.

A hipótese principal, neste caso, é a de que os termos selecionados dos extratos específicos tendem a ser mais estreitamente relacionados ao tópico da consulta do que os termos dos extratos genéricos. Porém, se os documentos forem irrelevantes à consulta, os termos selecionados também poderão empobrecer a consulta reformulada.

O fato de o GistSumm utilizar uma medida de similaridade (cosseno) diferente da medida de similaridade empregada no RFQBS (Dice) não implica qualquer problema: o GistSumm calcula a similaridade entre sentenças, enquanto os sistemas de recuperação, entre consulta e documento. Os resultados de recuperação podem variar se qualquer um dos dois adotarem uma medida de similaridade diferente. Essa variação não será investigada aqui.

5.3.3 RFQBM: RPR com extratos específicos multi-documentos

A diferença do RFQBM com relação aos demais sistemas com RPR é a adoção da SA multi-documentos, um recurso disponibilizado prontamente pelo GistSumm e, por esse motivo, incorporado a este trabalho. Para gerar extratos específicos multi-documentos, o GistSumm considera o conjunto de documentos como se fosse um único texto a sumarizar e, então, conduz a sumarização de modo equivalente ao descrito na seção anterior. Ao que parece não há trabalhos realizados anteriormente que façam uso de extratos de múltiplos documentos na RPR. Esse uso se justifica tendo em vista que nem todos os documentos pseudo-relevantes terão a mesma proporção de suas sentenças selecionadas para compor o extrato. Dessa forma, os extratos poderão conter somente sentenças que são fortemente relacionadas com a consulta inicial, fornecendo, então, bons termos para a reformulação da consulta.

5.3.4 RFFullDoc: RPR com documentos completos

O RFFullDoc incorpora os processos descritos na arquitetura da Figura 11, exceto a sumarização. Os termos que servem à RPR são selecionados, então, diretamente dos documentos, com base na medida rsv, como antes assinalado. Dessa forma, os resultados do

RFFullDoc servem para fazer comparação com os sistemas que usam extratos na RPR. A Figura 12 apresenta a arquitetura desse sistema.

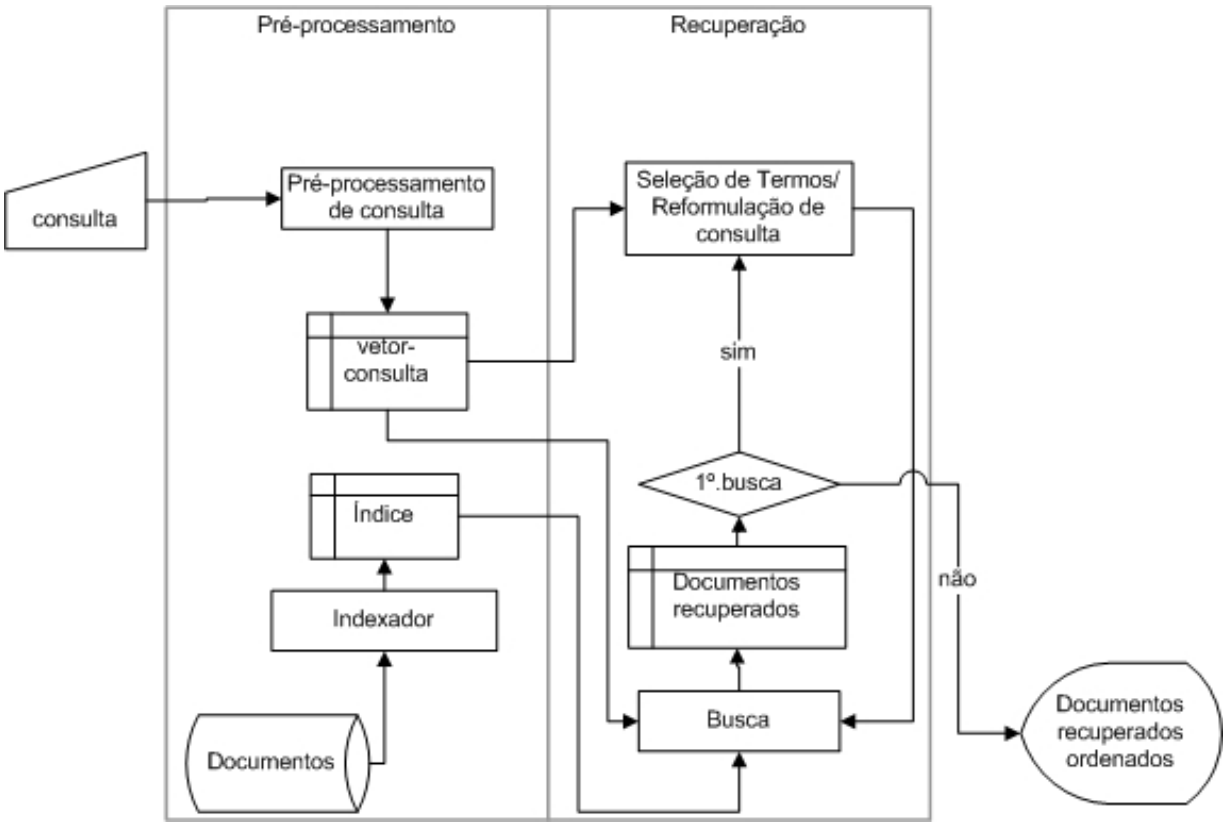


Figura 12 – Arquitetura do sistema RFFullDoc

O problema com o RFFullDoc é que ele poderá selecionar termos de documentos irrelevantes e de seções irrelevantes de documentos relevantes, fazendo com que termos ruins (ou seja, não relacionados a consulta) sejam inseridos na consulta reformulada.

Como se pode notar, os sistemas acima descritos e sintetizados na Tabela 6 lançam mão de técnicas usuais da RI, quer para indexação, quer para a RPR, a fim de proporcionar uma busca mais efetiva. No entanto, eles visam à recuperação de coleções particulares em português, constituindo, assim, exemplares interessantes e, em alguns casos, também genuínos, como é o caso do RFQBM. No próximo capítulo são apresentadas suas avaliações, reproduzindo-se a tarefa do CLEF 2004.

Tabela 6 – Sistemas propostos

Foco	Sistema	Descrição
Indexação	RExt	RI com extratos na indexação.
	RDocExt	RI com extratos e documentos na indexação.
Realimentação de pseudo-relevantes (RPR)	RFQBS	RI usando extratos específicos mono-documento na RPR.
	RFQBM	RI usando extratos específicos multi-documentos na RPR.
	RFGenS	RI usando extratos genéricos mono-documento na RPR.
	RFFullDoc	RI usando documentos na RPR.

6 Avaliações dos sistemas

A avaliação dos sistemas explorados neste trabalho tomou por base a tarefa *ad hoc* monolíngüe do CLEF 2004, para a língua portuguesa. Essa tarefa visou verificar a habilidade dos sistemas em recuperar documentos relevantes a um conjunto de consultas, contidos em uma coleção estática. Foi reproduzida integralmente a avaliação proposta no CLEF 2004 para os 23 sistemas inscritos. A diferença é que os sistemas aqui desenvolvidos não contribuíram para a formação do conjunto de documentos relevantes, como contribuíram os 23 sistemas inscritos na ocasião. Ou seja, os sistemas apresentados neste trabalho simulam essa tarefa sem ter participado dela de fato.

As próximas seções descrevem os dados e a metodologia de execução dos experimentos, assim como seus resultados.

6.1 Coleção de testes usada na avaliação

Para a avaliação dos sistemas, a coleção CHAVE (SANTOS; ROCHA, 2004) foi utilizada por dois motivos: por ser a coleção adotada no CLEF 2004 e por ser uma coleção de documentos escritos em português. Uma amostra de um documento da coleção é mostrada na Figura 13. Os documentos, 55.070 no total, são marcados em SGML e possuem informações tais como: número de identificação, data de publicação, categoria e o texto propriamente dito.

A coleção possui 50 tópicos (Anexo 1) marcados em SGML, com os quais devem ser elaboradas as consultas para buscas. Segundo o julgamento dos assessores do CLEF, um tópico para o qual o documento da Figura 13 é relevante é apresentado na Figura 14. Ele apresenta três campos: título (PT-title), descrição (PT-desc) e narrativa (PT-narr). O título é o principal descritor do tópico; a descrição consiste de uma sentença especificando a

necessidade de informação, enquanto que a narrativa indica o que o documento relevante deve conter.

A Figura 15 mostra parte do arquivo com os julgamentos de relevância para o tópico da Figura 14. Nesse arquivo, cada linha representa um documento julgado. As colunas representam, respectivamente:

i) o identificador do tópico;

ii) a segunda coluna é sempre 0 e não é usada para os fins de avaliação do CLEF;

iii) o identificador do documento;

iv) a relevância do documento identificado por (iii) com relação ao tópico identificado por (i), sendo que 0 indica irrelevância e 1 relevância.

```
<DOC>
<DOCNO>PUBLICO-19951005-038</DOCNO>
<DOCID>PUBLICO-19951005-038</DOCID>
<DATE>19951005</DATE>
<CATEGORY>Economia</CATEGORY>
<TEXT>
Decisão nas mãos do Governo alemão
Tribunal aprova extradição de Nick Leeson
O Tribunal de recurso da cidade alemã de Frankfurt deu ontem luz verde
para a extradição de Nick Leeson, o antigo quadro do banco Barings em
Singapura acusado de ter provocado a falência da instituição, para
aquela cidade asiática.
Leeson, 28 anos, é suspeito de estar na origem de perdas de 1,3 mil
milhões de dólares (mais de 195 milhões de contos) do Baring Brothers,
devido a apostas ruinosas no mercado de futuros de Tóquio e na bolsa
desta cidade.
O caso remonta a meados de Fevereiro passado, quando foram detectadas
oficialmente as perdas da instituição -- ao longo do processo, vários
jornais e entidades referiram que a direcção do Barings tinha
conhecimento da existência de elevados prejuízos mas que nada teria
feito para os corrigir, revelando algum laxismo no controlo interno.
Pouco depois do anúncio da falência do banco, que posteriormente foi
adquirido pelos holandeses do ING, as autoridades de Singapura
acusaram Nick Leeson de falsificação de papéis e de actividade
fraudulenta, tendo avançado de imediato com um pedido de extradição.
Ao todo, as autoridades de Singapura acusam Leeson de quatro
falsificações de documentos e de oito fraudes tanto no banco onde
trabalhava como na Bolsa de Singapura.
No Reino Unido, o gabinete de combate à fraude também levou a cabo um
inquérito, tendo apurado que Nick Leeson não cometera qualquer
irregularidade, o que impedia um pedido de extradição do jovem quadro
para o seu país de origem.
Já em Setembro, um tribunal da City londrina decidiu que havia motivos
para pedir a extradição de Leeson para o Reino Unido. A decisão foi
tomada na sequência de uma série de queixas de antigos accionistas do
Barings e de pessoas que faziam os seus investimentos através do
bicentenário banco britânico. Desde que foi preso que Leeson pede para
ser extraditado para o Reino Unido, alegando as condições desumanas
com que são tratados os presos em Singapura.
Mal se soube da decisão do Tribunal de Frankfurt, que deu como
provadas 11 das 12 acusações de Singapura, o advogado do corretor,
Eberhard Kempf, anunciou que irá recorrer para o Tribunal
Constitucional da Alemanha, a última entidade a quem poderá recorrer.
No entanto, após a decisão do Tribunal, o Governo alemão ainda tem uma
palavra a dizer, procedendo à extradição ou evitando-a.
</TEXT>
</DOC>
```

Figura 13 – Exemplo de documento da coleção CHAVE-2004.

```
<top>
<num> C202 </num>
<PT-title> Prisão de Nick Leeson </PT-title>
<PT-desc> Encontrar documentos sobre a prisão de Nick Leeson e as causas que o
levaram à cadeia. </PT-desc>
<PT-narr> Documentos relevantes devem relatar as razões da prisão de Nick Leeson e
sua subsequente prisão. </PT-narr>
</top>
```

Figura 14 – Exemplo de tópico da coleção CHAVE 2004

```
202 0 PUBLICO-19951005-038 1
202 0 PUBLICO-19951006-018 0
202 0 PUBLICO-19951011-036 0
202 0 PUBLICO-19951012-076 0
202 0 PUBLICO-19951013-140 0
202 0 PUBLICO-19951013-147 0
202 0 PUBLICO-19951019-043 0
202 0 PUBLICO-19951023-004 0
202 0 PUBLICO-19951026-030 0
202 0 PUBLICO-19951026-045 0
202 0 PUBLICO-19951028-047 0
202 0 PUBLICO-19951031-020 0
202 0 PUBLICO-19951031-064 0
202 0 PUBLICO-19951031-146 0
202 0 PUBLICO-19951114-136 0
202 0 PUBLICO-19951115-026 0
202 0 PUBLICO-19951125-030 1
```

Figura 15 – Registros de relevância para o tópico 202

Uma característica importante dessa coleção é o número de documentos relevantes por tópico: em média 13. Porém, para muitos tópicos (cerca de 22) há no máximo três documentos relevantes e para quatro tópicos não há nenhum documento relevante. Essa característica dificulta muito a recuperação, podendo fazer com que os sistemas avaliados apresentem precisões muito baixas.

6.1.1 Processamento dos dados da coleção CHAVE

Na execução dos sistemas construídos durante este trabalho, uma consulta foi formada automaticamente para cada tópico, com base no título e na descrição do tópico. Para a formação, ambos os campos passaram pelos processos de tokenização, remoção de *stopwords* e remoção de palavras irrelevantes, antes descritos, para a formação da representação interna

da consulta (vetor-consulta). As palavras consideradas irrelevantes são aquelas comuns a muitos tópicos e que, portanto, não podem ser consideradas palavras-chave para a busca. Em geral, não são palavras de domínio específico, mas palavras pertencentes a frases que remetem à tarefa de busca, como ilustra a Figura 16.

algumas, descrevem, descrevendo, detalhar, discussões, discutindo, documentos, encontrar, encontre, falando, fornecendo, informação, informações, particular, relatam relatando, relatórios

Figura 16 – Lista de palavras irrelevantes dos tópicos

Para o tópico 202, por exemplo, os seguintes termos formaram a consulta: prisa (3.09), nick (10.75), lesson (13.54), caus (1.60), lev (1.38), cade (3.26). O valor entre parênteses indica o peso de cada termo na consulta, resultante do produto entre a freqüência do termo na consulta e a medida idf do termo (Eq. 8, p. 35). Neste exemplo, os termos 'nick' e 'leeson' têm pesos maiores, pois são mais raros na coleção que os demais.

As consultas formadas com base na descrição e no título dos tópicos são chamadas de consultas longas. Consultas curtas são formadas somente com base no título. Nos experimentos com extratos, apenas foram consideradas consultas longas, pois os resultados com consultas curtas no RDoc faziam com que a recuperação inicial retornasse poucos documentos relevantes dentre os cinco primeiros. Logo, isso poderia prejudicar a RPR devido à grande quantidade de documentos irrelevantes dentre os pseudo-relevantes.

Trabalhos como o de Llopis et al. (2004) e Adafre et al. (2004) também utilizam consultas longas em seus experimentos com a CHAVE 2004. Além disso, o CLEF estimula a criação de consultas longas automaticamente para efeito de comparação entre os sistemas, embora não restrinja a consulta a esse único tipo. Uma vez criadas, as 50 consultas foram submetidas aos sistemas propostos. Cada sistema, então, gerou a lista de documentos com maior grau de similaridade com as respectivas consultas (*hitlist*). O tamanho da *hitlist* não foi

limitado por um valor mínimo de similaridade, mas pela quantidade: 300 documentos foram considerados para o experimento.

Os sistemas geram a lista de documentos recuperados no formato de entrada do programa trec_eval[18], que é um programa utilizado pelo CLEF e TREC para avaliação da efetividade dos sistemas. Particularmente neste trabalho, são exibidas somente as medidas P_k, para k igual a 5, 10, 15 e 20, *R-Precision*, MAP e a precisão interpolada para os 11 graus padrões de revocação, todas apresentadas no Capítulo 2.

A Figura 17 mostra parte da lista de saída do sistema RDoc para o tópico 202. As linhas representam cada documento recuperado. As colunas representam:

i) o identificador do tópico;

ii) o número da consulta criada para esse tópico, sendo que esse campo não é utilizado pelo trec_eval e é sempre igual a Q0;

iii) o identificador do documento;

iv) a posição do documento na lista de documentos recuperados;

v) o grau de similaridade do documento com a consulta;

vi) o identificador do sistema.

[18] O programa trec_eval pode ser obtido em http://trec.nist.gov/results.html, último acesso em 13/06/2006.

```
202 Q0 PUBLICO-19950304-040 0 0.6555 RDoc
202 Q0 PUBLICO-19951207-076 1 0.6492 RDoc
202 Q0 PUBLICO-19951125-030 2 0.6422 RDoc
202 Q0 PUBLICO-19950303-034 3 0.6308 RDoc
202 Q0 PUBLICO-19950720-037 4 0.6060 RDoc
202 Q0 PUBLICO-19950309-041 5 0.5969 RDoc
202 Q0 PUBLICO-19951005-038 6 0.5936 RDoc
202 Q0 PUBLICO-19951023-004 7 0.5931 RDoc
202 Q0 PUBLICO-19950305-029 8 0.5924 RDoc
202 Q0 PUBLICO-19951006-018 9 0.5886 RDoc
202 Q0 PUBLICO-19950302-036 10 0.5868 RDoc
202 Q0 PUBLICO-19950304-128 11 0.5864 RDoc
202 Q0 PUBLICO-19950804-040 12 0.5824 RDoc
202 Q0 PUBLICO-19950728-174 13 0.5816 RDoc
202 Q0 PUBLICO-19950328-040 14 0.5798 RDoc
202 Q0 PUBLICO-19950330-044 15 0.5740 RDoc
202 Q0 PUBLICO-19950308-027 16 0.5738 RDoc
202 Q0 PUBLICO-19950519-043 17 0.5734 RDoc
202 Q0 PUBLICO-19950728-137 18 0.5716 RDoc
202 Q0 PUBLICO-19950503-037 19 0.5708 RDoc
202 Q0 PUBLICO-19950829-026 20 0.5695 RDoc
202 Q0 PUBLICO-19950301-019 21 0.5689 RDoc
202 Q0 PUBLICO-19950228-130 22 0.5675 RDoc
202 Q0 PUBLICO-19950428-043 23 0.5659 RDoc
202 Q0 PUBLICO-19950714-034 24 0.5555 RDoc
202 Q0 PUBLICO-19950906-034 25 0.5540 RDoc
202 Q0 PUBLICO-19950505-035 26 0.5529 RDoc
202 Q0 PUBLICO-19950314-035 27 0.5526 RDoc
202 Q0 PUBLICO-19950307-040 28 0.5494 RDoc
```

Figura 17 – Exemplo de saída do sistema RDoc para o tópico 202

6.2 Síntese dos resultados

Os resultados da avaliação com a coleção CHAVE 2004 são apresentados nesta seção, juntamente com as médias de desempenho dos 23 sistemas que participaram do CLEF 2004 na tarefa *ad hoc* para o português. Todos os sistemas foram comparados com o *baseline*, o RDoc. A diferença relativa entre os sistemas e o *baseline* também é exibida. Com a finalidade de verificar se as diferenças entre os sistemas propostos e o RDoc são estatisticamente significantes, o teste t emparelhado e o teste dos sinais (discutidos no Capítulo 2) também foram realizados com nível de significância (α) igual a 0.05. Esse nível indica a probabilidade de rejeitar a hipótese nula quando ela é verdadeira (vide Seção 2.2.2). Nas tabelas, o símbolo * ao lado da diferença relativa indica que há evidências suficientes para determinar que a

diferença é estatisticamente significante para o teste dos sinais. A ausência do símbolo indica que, para o teste dos sinais, os métodos de recuperação são equivalentes em relação à performance, ou seja, a diferença não é estatisticamente significante (hipótese nula). Já o símbolo $^{\#}$ indica que a diferença é estatisticamente significante, segundo o teste t emparelhado, e sua ausência indica a hipótese nula.

Na seção que segue, são apresentados os resultados que incentivaram o uso de consultas longas e da medida de ponderação adotada. Após essa seção, serão mostrados os resultados dos sistemas que utilizam extratos no seu processamento.

6.2.1 Escolha por consultas longas e pela tf normalizada através de logaritmo

Nos experimentos com os sistemas desenvolvidos neste trabalho, foi feita a opção por consultas longas e também pela ponderação de documentos usando a normalização logarítmica (Eq. 12, p. 37). Essas duas escolhas foram motivadas pelo desempenho do *baseline* com essas configurações.

As tabelas 7 e 8 e o gráfico da Figura 18 apresentam as precisões do RDoc usando consultas longas e curtas. A Tabela 7 exibe as precisões para os primeiros 5, 10, 15 e 20 documentos recuperados, respectivamente, P5, P10, P15 e P20, enquanto a Tabela 8 exibe as medidas MAP e *R-Precision*. A diferença entre o RDoc com consultas longas e com curtas é mostrada entre parênteses. O teste t somente não indicou como significativa a diferença com P15, já o teste dos sinais somente indicou como significativas as diferenças com P5 e *R-Precision*.

Tabela 7 – Precisões P5, P10, P15, P20 para o Rdoc usando consultas curtas e longas

Sistemas (tipo de consulta)	P5	P10	P15	P20
RDoc (consultas longas)	0.3560	0.2660	0.2240	0.2000
RDoc (consultas curtas)	0.2600 (-27.0%)* #	0.2120 (-20.3%)#	0.1840 (-17.8%)	0.1530 (-23.5%)#

Tabela 8 – MAP e R-Precision para o RDoc usando consultas curtas e longas

Sistema (tipo de consulta)	MAP	R-Precision
RDoc (consultas longas)	0.4203	0.3923
RDoc (consultas curtas)	0.3125 (-25.6%)	0.2784 (-29.0%)* #

Esses resultados mostram o quão difícil foi recuperar documentos relevantes da coleção CHAVE com consultas curtas. Como a primeira recuperação dos sistemas que usam RPR é feita da mesma forma que o RDoc, usando consultas curtas ter-se-ia muito poucos documentos relevantes dentre os primeiros 5 recuperados (P5). Quanto menor a P5, menor o número de documentos relevantes dentre os pseudo-relevantes. Dessa forma, a recuperação com consultas curtas poderia fazer que, na etapa de RPR, termos ruins (ou seja, não relacionados ao tópico da consulta) fossem acrescentados à consulta reformulada.

O gráfico da Figura 18 mostra a diferença de performance entre o RDoc usando consultas curtas e o mesmo sistema usando consultas longas. Para todos os 11 graus padrões de revocação, o RDoc usando consultas longas apresentou uma proporção maior de documentos relevantes dentre os recuperados.

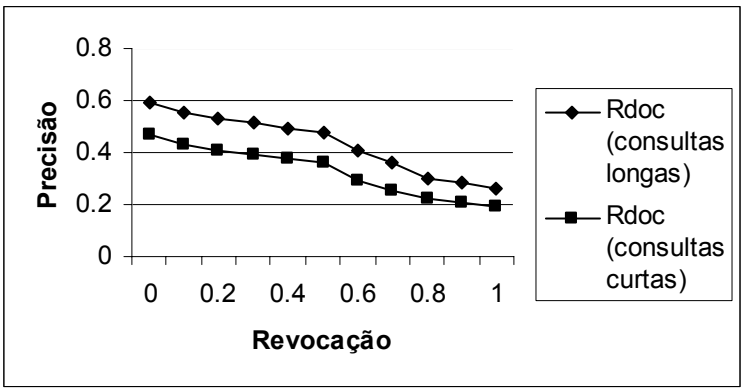


Figura 18 – Precisão para os 11 graus de revocação (interpolada) para consultas curtas e longas

Já no que se refere à ponderação de termos, inicialmente foi considerada a utilização da medida tf-idf, com tf normalizada com a freqüência do termo mais recorrente no documento que é, segundo Baeza-Yates e Ribeiro-Neto (1999), o esquema mais conhecido de ponderação de termos. Os resultados mostrados nas tabelas 9 e 10 e na Figura 19 fizeram com que a medida da Eq. 12 (p. 37) fosse escolhida. Somente o RDoc foi testado nos experimentos. Para esses resultados, o teste t apontou as diferenças com P5, P20 e *R-Precision* como estatisticamente significantes, já o teste dos sinais apenas apontou como significativas as diferenças entre P5 e *R-Precision*.

Tabela 9 – Precisões P5, P10, P15 e P20 para o RDoc usando diferentes formas de ponderação

Sistema	P5	P10	P15	P20
RDoc (Eq. 12)	0.3560	0.2660	0.2240	0.2000
RDoc (tf-idf)	0.2960 (-16.8%)$^{*\#}$	0.2320 (-12.8%)	0.2013 (-10.1%)	0.1760 (-12.0%)$^{\#}$

Tabela 10 – MAP e R-Precision para o RDoc usando diferentes formas de ponderação

Sistema	MAP	R-Precision
RDoc (Eq. 12)	0.4203	0.3923
RDoc (tf-idf)	0.3190 (-24.1%)	0.2812 (-28.3%)$^{*\#}$

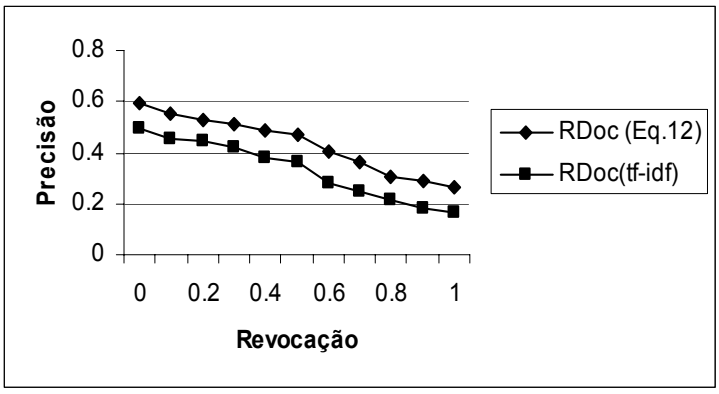


Figura 19 – Precisão para os 11 graus de revocação (interpolada) para diferentes formas de ponderação

É visível a diferença dos resultados. O RDoc utilizando tf-idf teve desempenho inferior ao RDoc que usa normalização logarítmica (Eq. 12) em todas as medidas. Esses

resultados apontam que não usar a normalização baseada na freqüência máxima do documento não prejudica os resultados.

O resultado inferior da medida P5 do sistema RDoc usando tf-idf indica que ele fornece menos documentos relevantes dentre os primeiros cinco que método usando normalização logarítmica, o que prejudicaria a recuperação dos sistemas que utilizam RPR. Por esse motivo, a medida usando a normalização logarítmica foi escolhida para todos os sistemas nos demais experimentos.

6.2.2 Avaliação de extratos na indexação

Os resultados dos experimentos com as versões de sistemas RExt80 e RExt60, e do sistema RDocExt são apresentados nesta subseção, juntamente com os resultados do RDoc que utiliza somente os documentos na indexação. Os números 80 e 60 indicam as taxas de compressão dos extratos gerados pelo GistSumm.

A Tabela 11 exibe as precisões para os primeiros 5, 10, 15 e 20 documentos recuperados, respectivamente, P5, P10, P15 e P20. Entre parênteses são apresentadas as diferenças relativas entre as precisões dos sistemas e as precisões do RDoc. De acordo com o teste t de significância, todas as diferenças, excetuando-se a diferença da Média CLEF 2004, são estatisticamente significantes. O teste dos sinais somente não indicou que as diferenças entre a RExt60 e o RDoc são significativas, indicando que o desempenho do RDoc não foi superior ao desempenho apresentado pela RExt60 mais freqüentemente que a média esperada.

Tabela 11 – Precisões P5, P10, P15, P20 para diversas formas de indexação

Sistemas	P5	P10	P15	P20
RDoc	0.3560	0.2660	0.2240	0.2000
RExt80	0.2400 (-33%)$^{*\#}$	0.2040 (-23%)$^{*\#}$	0.1667 (-26%)$^{*\#}$	0.1460 (-27%)$^{*\#}$
RExt60	0.2840 (-20%)$^{\#}$	0.2280 (-14.3%)$^{\#}$	0.2000 (-10.7%)$^{\#}$	0.1720 (-14.0%)$^{\#}$
RDocExt	0.2840 (-20%)$^{*\#}$	0.2160 (-19%)$^{*\#}$	0.1747 (-22%)$^{*\#}$	0.1440 (-28%)$^{*\#}$
Média CLEF 2004	0.3550 (-0.3%)	0.2726 (+2.5%)	0.2240 (0.0%)	0.2310 (+15%)

Esses resultados mostram que a precisão diminui à medida que o número de documentos considerados no cálculo da precisão aumenta. Isso é comum, já que há poucos documentos relevantes na coleção e os sistemas tendem a apresentar os documentos relevantes nas primeiras posições. Eles também mostram que as precisões do RDoc são superiores a todas as demais, indicando que a recuperação utilizando extratos não tem desempenho próximo à recuperação usando o documento. Também é possível verificar que as precisões da RExt80 são inferiores às precisões apresentadas pela RExt60 para todas as medidas, indicando que os valores de precisão são inversamente proporcionais à taxa de compressão dos extratos utilizados na indexação. O desempenho do RDocExt é muito próximo da RExt80, mostrando que o fato de usar tanto extratos quanto documentos na indexação não é suficiente para que suas precisões serem superiores às precisões apresentadas pelo RDoc.

A Tabela 12, para os mesmos sistemas, apresenta as medidas MAP e *R-precision*. Assim como na Tabela 11, a porcentagem indica a diferença relativa entre os sistemas e o *baseline*. Os resultados dessa tabela levam à mesma constatação sobre os resultados apontados na Tabela 11: os extratos não são úteis na indexação. Os dois testes de significância indicam que as diferenças de *R-precision* são significantes.

Tabela 12 – MAP e R-Precision para as diversas formas de indexação

Sistemas	MAP	R-Precision
RDoc	0.4203	0.3923
RExt80	0.2342 (-44%)	0.2215 (-43%)* #
RExt60	0.2993 (-28.8%)	0.2834 (-27.7%)* #
RDocExt	0.2645 (-37%)	0.2629 (-33%)* #
Média CLEF 2004	-------	0.3428 (-13%)

A Figura 20 mostra o gráfico da precisão interpolada para os 11 graus padrões de revocação que os sistemas apresentaram. A precisão interpolada do RDoc para os 11 graus é superior às apresentadas pelos demais sistemas, mostrando que, para as mesmas quantidades de documentos relevantes da coleção recuperados, é descartado um número maior de documentos irrelevantes.

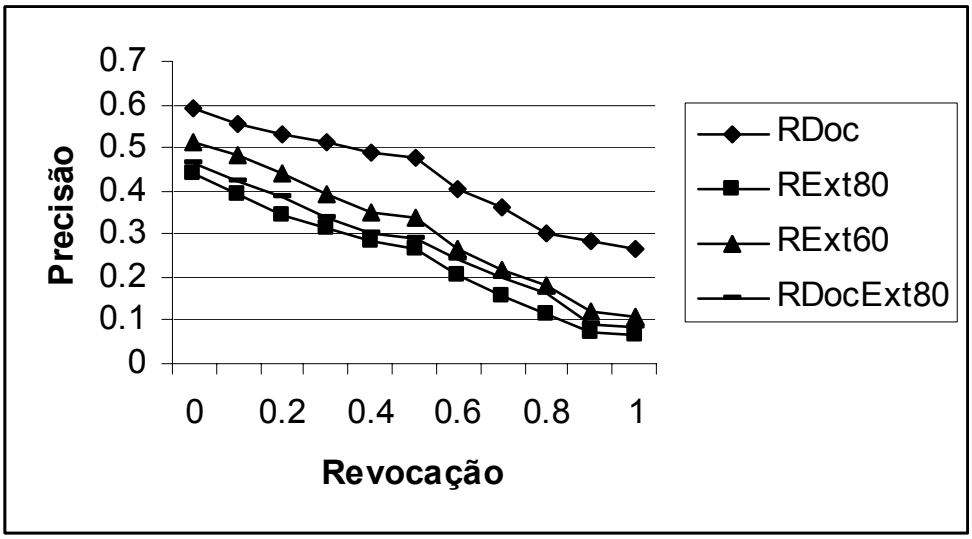


Figura 20 – Precisão para os 11 graus de revocação (interpolada) para formas distintas de indexação

O desempenho dos sistemas que usam extratos na indexação pode ser justificado pelo fato de os extratos genéricos extraírem somente as sentenças relacionadas ao tópico principal (ou seja, o mais enfatizado no documento), o que pode fazer com que não sejam recuperados documentos cuja informação relevante para o tópico não seja a mais enfatizada. Sormunen (2002) constatou que, para 38 tópicos, 36% dos documentos da coleção TREC 7 e TREC 8

julgados relevantes não tinham as informações relevantes como as mais enfatizadas. O fato de os documentos poderem ser considerados relevantes tanto na TREC como no CLEF, mesmo que a informação relevante não seja o tópico principal do documento, faz que os extratos genéricos se tornem pouco efetivos na recuperação, quando utilizados na tarefa de indexação. Por exemplo, há cinco documentos julgados relevantes no CLEF 2004 para o tópico 218, que aparecem dentre os 30 primeiros documentos recuperados pelo sistema RDoc. No entanto, esses documentos não foram recuperados pela RExt80. O motivo pode ser constatado pela leitura dos dois documentos, apresentados abaixo, juntamente com os respectivos extratos gerados pelo GistSumm. Esses documentos foram considerados relevantes para o tópico 218 (Figura 21), segundo o julgamento realizado no CLEF.

<top>
<num> C218 </num>
<PT-title> Andreotti e a Mafia </PT-title>
<PT-desc> Encontrar informação sobre acusações judiciais apresentadas contra o ex-primeiro-ministro italiano, Giulio Andreotti, acusando-o de pertencer ou de estar associado à Mafia. </PT-desc>
<PT-narr> Documentos relevantes devem detalhar as acções legais; a mera listagem das acusações sem informação adicional não é suficiente </PT-narr>
</top>

Figura 21 – Tópico 218 da coleção CHAVE 2004

Documento: PUBLICO-19950312-038

<DOC>
<DOCNO>PUBLICO-19950312-038</DOCNO>
<DOCID>PUBLICO-19950312-038</DOCID>
<DATE>19950312</DATE>
<CATEGORY>Mundo</CATEGORY>
<TEXT>
Clima de tensão na Sicília
Mafia em guerra contra «arrependidos»
Eduardo Tessler em Roma
Os «padrinhos» sicilianos voltaram a atacar para lembrar que, mesmo com os seus principais chefes presos, a Mafia não desapareceu. A guerra agora é contra os «arrependidos» que aceitaram colaborar com a Justiça. É verdade que estes estão ultra-protegidos, mas os seus familiares não e, em apenas nove dias, oito foram já assassinados.
A Mafia não se fazia tão viva desde os assassínios dos juízes Giovanni Falcone e Paolo Borsellino, em 1992. O início de Março, porém, serviu para

que ninguém esqueça do potencial da mais perigosa organização criminosa do mundo, a Cosa Nostra siciliana. E não importa que os poderosos chefes Salvatore Riina, Michele Greco, Nitto Santapaola, Francesco Madonia, Pippo Caló e Bernardo Brusca estejam presos.

A estratégia do terror e sangue da Mafia renasceu das cinzas e em apenas nove dias deixou oito mortos só na região de Palermo, com uma coincidência peculiar: todas as vítimas têm algum tipo de relação com os mafiosos «arrependidos», aqueles que depois de presos passam a colaborar com a Justiça em troca da redução das penas.

Para complicar ainda mais o enredo do que parece ser o início de uma nova «Guerra de Mafia», o marechal dos carabineiros Antonino Lombardo suicidou-se e deixou uma carta de despedida cheia de enigmas indecifráveis. O maior de todos: qual o motivo que o levou a tal gesto extremo?

Dois dias antes, o ex-primeiro-ministro por sete vezes Giulio Andreotti era acusado oficialmente de associação mafiosa, com início de processo marcado para Setembro. Para responder aos que o consideram um mafioso, o senador vitalício lançou sexta-feira o livro «Eu e a Mafia», a dar a sua versão aos factos. Andreotti desmente, por exemplo, o famoso beijo que teria dado em Totó Riina.

«Não tenho dúvidas de que se trata de uma guerra de Mafia", avalia o vice-presidente da Câmara e ex-presidente da comissão antimafia do Parlamento Luciano Violante. «É uma guerra entre as famílias, entre grupos, para apresentar a maior capacidade ofensiva e de intimidação". Talvez o mais importante seja que uma nova geração de mafiosos anuncia que chegou para ficar. Uma organização criminosa chefiada por Bernardo Provenzano, inimigo pessoal de Totó Riina.

O perigo de se chamar Buscetta

Os sinais são evidentes. A Mafia eliminou o sobrinho de Tommaso Buscetta, o genro do ex-chefe Pippo Mineo, o primo de Salvatore Contorno e ainda o informador do marechal Lombardo. É aquilo a que a polícia chama «vingança transversal». Sem poder aproximar-se dos super-protegidos «arrependidos», a Mafia ataca os seus parentes e amigos mais próximos. Foi uma técnica muito usada no final de 1984, quando a Cosa Nostra, já sob o comando de Riina, assassinou quatro parentes de ex-mafiosos. Um deles era o cunhado de Buscetta, Pietro Buscetta. «Don Masino», como era chamado, coleccionava assim mais uma baixa na família. Antes a Mafia já liquidara dois filhos, outro cunhado, um irmão, um sobrinho, um genro e dois colaboradores. O perigo de ser Buscetta é tanto que naquela época a sua irmã Serafina disse «odiá-lo como irmão». Na semana passada, o irmão de Domenico Buscetta, o sobrinho assassinado do «arrependido», pediu: «Se o meu tio é um homem de verdade, deve suicidar-se, para salvar o que resta da família».

Há quem diga que Tommaso Buscetta provocou os seus ex-companheiros. No dia do assassinato do sobrinho, tinha publicado um comentário no jornal «La Repubblica» a propósito da série de televisão «La Piovra», sobre a Mafia. Don Masino não poupou comparações: «A realidade é pior do que a ficção».

A prisão é mais segura para Don Tano

«Não tenho dúvidas de que o objectivo da Cosa Nostra com os novos atentados é intimidar os arrependidos que colaboram com a Justiça", avalia o procurador da República de Caltanissetta, Giovanni Tinebra. E a sua teoria parece confirmar-se com o avançar dos tempos. O mafioso Gaetano Badalamenti, ex-aliado e hoje inimigo de Tommaso Buscetta, estava a anunciar sua colaboração com a Justica. Mas mudou de ideias e avisou que não pretende ser ouvido pelos juízes.

Badalamenti, 71 anos, foi o chefe da Cosa Nostra no início dos anos 80. Mas o seu cedro foi tomado ao ritmo de atentados por Totó Riina. «Don Tano», como era conhecido, fugiu para os Estados Unidos e depois para o Brasil, onde manteve contatos com Buscetta. Lá coordenava o tráfico internacional de drogas, em comum acordo com a família Gambino, de Nova Iorque. Usando o nome falso de Paulo Ares Barbosa, Badalamenti foi preso em Madrid, em 1983. Hoje cumpre uma pena de 45 anos na prisão de Fairton, Estados Unidos, perto dos familiares, que vivem em Nova Jersey.

Segundo Tommaso Buscetta, Badalamenti participou no assassínio do jornalista Mino Pecorelli -- um dos mistérios inexplicáveis da Itália -- e sabe muitos detalhes do sequestro e morte de Aldo Moro e do atentado contra o ex-chefe de polícia da Sicília, Carlo Alberto dalla Chiesa. Mas depois do clima de guerra que tomou conta da Sicília, Don Tano decidiu permanecer na sua cela americana.
A tensão na Sicília é tão grande que até mesmo uma pacífica marcha em favor da mulher, dia 8 de Marco, teve pouca adesão em Corleone, sede histórica da Cosa Nostra. As mulheres da cidade têm outra vez medo de sair à rua. «A Mafia não é apenas aquela organização criminosa que conhecemos", observa o cardeal arcebispo de Palermo, Salvatore Pappalardo. «Mafia é também o comportamento da administração pública, dos políticos, o uso dos meios de comunicação. O mal não teria tanta força se encontrasse resistência social que dificultasse as suas acções». O cardeal condena o silêncio dos sicilianos em casos de atentados mafiosos e pede uma maior participação popular. «É a única maneira de enfrentar os criminosos», alega.
De qualquer forma, a luta do Estado italiano contra a Mafia ainda está longe de terminar em sucesso. A polícia federal enviou novos agentes para garantir a calma na ilha da Sicília, embora a raiz do problema não seja desafiada de frente. Uma das raras vozes optimistas nesta guerra vem do juiz Antonino Caponnetto, professor e mentor judicial de Falcone e Borsellino. «Eu já estou velho e certamente não verei a derrota definitiva da Mafia», admitiu. «Mas os mais jovens, os estudantes de hoje, com certeza viverão numa Itália sem Mafia. Disso tenho certeza». Se depender do clima de tensão que envolve a Sicília hoje, a derrota definitiva da Mafia ainda vai demorar.
</TEXT>
</DOC>

A prisão é mais segura para Don Tano «Não tenho dúvidas de que o objectivo da Cosa Nostra com os novos atentados é intimidar os arrependidos que colaboram com a Justiça", avalia o procurador da República de Caltanissetta, Giovanni Tinebra. E a sua teoria parece confirmar se com o avançar dos tempos. O mafioso Gaetano Badalamenti, ex aliado e hoje inimigo de Tommaso Buscetta, estava a anunciar sua colaboração com a Justica. Mas mudou de ideias e avisou que não pretende ser ouvido pelos juízes. Badalamenti, 71 anos, foi o chefe da Cosa Nostra no início dos anos 80. Mas o seu cedro foi tomado ao ritmo de atentados por Totó Riina. «Don Tano», como era conhecido, fugiu para os Estados Unidos e depois para o Brasil, onde manteve contatos com Buscetta. Lá coordenava o tráfico internacional de drogas, em comum acordo com a família Gambino, de Nova Iorque. Usando o nome falso de Paulo Ares Barbosa, Badalamenti foi preso em Madrid, em 1983. Hoje cumpre uma pena de 45 anos na prisão de Fairton, Estados Unidos, perto dos familiares, que vivem em Nova Jersey. Segundo Tommaso Buscetta, Badalamenti participou no assassínio do jornalista Mino Pecorelli um dos mistérios inexplicáveis da Itália e sabe muitos detalhes do sequestro e morte de Aldo Moro e do atentado contra o ex chefe de polícia da Sicília, Carlo Alberto dalla Chiesa. Mas depois do clima de guerra que tomou conta da Sicília, Don Tano decidiu permanecer na sua cela americana. A tensão na Sicília é tão grande que até mesmo uma pacífica marcha em favor da mulher, dia 8 de Marco, teve pouca adesão em Corleone, sede histórica da Cosa Nostra. As mulheres da cidade têm outra vez medo de sair à rua. «A Mafia não é apenas aquela organização criminosa que conhecemos", observa o cardeal arcebispo de Palermo, Salvatore Pappalardo.

Figura 22 – Extrato para o documento PUBLICO-19950312-038 – com 80% de taxa de compressão

Documento PUBLICO-19950420-067

<DOC>
<DOCNO>PUBLICO-19950420-067</DOCNO>
<DOCID>PUBLICO-19950420-067</DOCID>
<DATE>19950420</DATE>
<CATEGORY>Mundo</CATEGORY>
<TEXT>
Aviões líbios partem
O Comité das sanções da ONU autorizou ontem os voos entre a Líbia e a Arábia Saudita para permitir que cerca de seis mil peregrinos se desloquem a Meca, anunciou o presidente do comité, Karel Kovanda. O pedido -- que prevê uma excepção no embargo aéreo imposto a Tripoli em 1992 -- foi feito pelo Egipto, cuja companhia aérea transportará a maior parte dos peregrinos líbios. Pouco antes da ONU ter dado a autorização formal, um avião líbio terá partido do aeroporto de Tripoli, em direcção à Arábia Saudita, mas o comité não quis discutir esta informação. Segundo noticiou a televisão líbia -- que transmitiu em directo a descolagem -- este aparelho, com 150 pessoas a bordo, partiu às 13h07 TMG, ainda sem autorização, e portanto em violação do embargo.
«Arrependido» arrepende-se
Um dos mais importantes arrependidos da Mafia, Francesco Marino Mannoia, anunciou ontem a intenção de não colaborar mais com a Justiça italiana, numa altura em que o seu testemunho era aguardado no processo do antigo primeiro-ministro Giulio Andreotti, acusado de cumplicidade com a Cosa Nostra. O antigo «químico» mafioso, especialista em drogas, acusou indirectamente o Estado italiano de abandonar os «arrependidos». No passado, Mannoia afirmou ter visto Andreotti em companhia de conhecidos «padrinhos», em dois encontros secretos na Sicília. O processo do ex-chefe de Governo deverá começar no dia 26 de Setembro, em Palermo. A televisão italiana considerou as declarações de Mannoia como um «protesto espectacular» e admitiu que outros «arrependidos» possam seguir o seu exemplo.
Refugiados sem alimentos
Após uma situação de pânico que provocou terça-feira dez mortos no campo de refugiados de Kibeho (sudoeste do Ruanda), os cerca de 130 mil ocupantes deste campo encontravam-se ontem refugiados numa colina e privados e água e alimentos, enquanto o governo os tentava convencer a regressar às suas tendas. Um porta-voz do Alto Comissariado da ONU para os Refugiados revelou que os militares estavam a cercar o campo, impedindo as organizações humanitárias de entrar no local. O objectivo das forças militares seria de «filtrar» as partidas, para identificar eventuais «criminosos» escondidos entre os refugiados. À semelhança dos campos instalados em países vizinhos, julga-se que muitos dos presumíveis autores dos massacres efectuados entre Abril e Julho do ano passado no Ruanda -- que provocaram cerca de 500 mil mortos, sobretudo tutsis e opositores hutus --, se introduziram nos campos de refugiados ruandeses.
</TEXT>
</DOC>
<DOC>

Refugiados sem alimentos Após uma situação de pânico que provocou terça feira dez mortos no campo de refugiados de Kibeho sudoeste do Ruanda), os cerca de 130 mil ocupantes deste campo encontravam se ontem refugiados numa colina e privados e água e alimentos, enquanto o governo os tentava convencer a regressar às suas tendas.

Figura 23 – Extrato para o documento PUBLICO-199550420-067 – com 80% de taxa de compressão

Os extratos do GistSumm mostrados nas figuras 22 e 23 claramente deixam de incluir informações relevantes ao respectivo tópico. Possivelmente porque, no primeiro caso (PUBLICO-19950312-038), o tópico apresentado na Figura 21 não é a informação mais enfatizada no documento; no segundo caso (PUBLICO-19950420-067), porque o documento é constituído por diferentes matérias jornalísticas com temas muito distintos. Cabe ressaltar que documentos com múltiplos tópicos e informação relevante superficial existem em vários ambientes, principalmente na Web (vide, por exemplo, os portais da UOL, <www.uol.com.br>, e Terra, <www.terra.com.br>). Isto certamente dificulta o desempenho do GistSumm.

Outro fato observado foi a existência de um grande número de documentos com o mesmo grau de similaridade para um mesmo tópico na versão RExt80. Para alguns tópicos, havia até seis documentos com a mesma similaridade. Dessa forma, alguns documentos irrelevantes apareceram numa posição acima dos relevantes *hitlist*. Isto ocorreu porque, em casos de empate, o documento aparece na *hitlist* na mesma seqüência de indexação.

Os resultados apontados acima não levam a conclusões distintas das obtidas por Sakai e Sparck-Jones (2001) que, embora tenham usado métodos de sumarização diferentes do GistSumm, constataram que extratos não são mais efetivos que documentos na tarefa de indexação.

6.2.3 Extratos na Realimentação de Pseudo-relevantes

Nesta seção são apresentados os resultados obtidos com os sistemas que fazem uso de Realimentação de Pseudo-relevantes, juntamente com suas comparações com o *baseline* RDoc. As diferenças entre eles (Tabela 13) não foram tão expressivas quanto as que foram apresentadas anteriormente: os testes de significância mostram que somente a diferença de P10 entre o RDoc e o RFQBM é significativa, sendo este o que apresentou o melhor desempenho.

Tabela 13 – Precisões P5, P10, P15, P20 para as diversas formas de RPR

Sistemas	P5	P10	P15	P20
RDoc	0.3560	0.2660	0.2240	0.2000
RFGenS	0.3480 (-2.2%)	0.2880 (+8.3%)	0.2373 (+5.9%)	0.2010 (+0.5%)
RFQBS	0.3600 (+1.1%)	0.2840 (+6.8%)	0.2360 (+5.3%)	0.2030 (+1.5%)
RFFullDoc	0.3551 (-0.2%)	0.2755 (+3.6%)	0.2259 (+0.8%)	0.1949 (-2.5%)
RFQBM	0.3720 (+4.5%)	0.2940 (+10.5%)* #	0.2307 (+3.0%)	0.2010 (+0.5%)
Média CLEF 2004	0.3550 (-0.3%)	0.2726 (+2.5%)	0.2240 (0.0%)	0.2310 (+15%)

De um modo geral, o desempenho dos sistemas com RPR piora à medida que o número de documentos considerado para o cálculo da precisão aumenta. Para P5, os sistemas cuja RPR se baseia em extratos específicos (QB) superam o *baseline* (Tabela 14) e, para P10 e P15, todos os sistemas com RPR o superam, assim como superam também a média de desempenho obtida pelos sistemas que participaram do CLEF 2004 (tabelas 15 e 16). Para P20, a maioria deles supera o *baseline*, mas fica abaixo da média do CLEF (Tabela 17).

Uma justificativa para que a precisão dos sistemas que usam extratos na RPR diminua à medida que um número maior de documentos é considerado, pode ser a adição de termos muito específicos a um ou dois documentos pseudo-relevantes à consulta durante a reformulação. Esses termos prejudicam a recuperação. Por exemplo, para o tópico 238, sobre

declarações da princesa Diana, o termo ‘bbc’ foi acrescentado à consulta pelo RFQBM, porque um dos documentos relatava uma entrevista da princesa à emissora BBC. Assim, documentos que não contêm esse termo, mas que são relevantes, tem seu grau de similaridade reduzido.

Tabela 14 – Comparação para P5

Sistemas	P5
RFQBM	**0.3720 (+4.5%)**
RFQBS	0.3600 (+1.1%)
RDoc	0.3560
RFFullDoc	0.3551 (-0.2%)
Média CLEF 2004	0.3550 (-0.3%)
RFGenS	0.3480 (-2.2%)

Tabela 15 – Comparação para P10

Sistemas	P10
RFQBM	**0.2940 (+10.5%)**
RFGenS	0.2880 (+8.3%)
RFQBS	0.2840 (+6.8%)
RFFullDoc	0.2755 (+3.6%)
Média CLEF 2004	0.2726 (+2.5%)
RDoc	0.2660

Tabela 16 – Comparação para P15

Sistemas	**P15**
RFGenS	0.2373 (+5.9%)
RFQBS	0.2360 (+5.3%)
RFQBM	**0.2307 (+3.0%)**
RFFullDoc	0.2259 (+0.8%)
Média CLEF 2004	0.2240 (0.0%)
RDoc	0.2240

Tabela 17 – Comparação para P20

Sistemas	**P20**
Média CLEF 2004	0.2310 (+15%)
RFQBS	0.2030 (+1.5%)
RFGenS	0.2010 (+0.5%)
RFQBM	**0.2010 (+0.5%)**
RDoc	0.2000
RFFullDoc	0.1949 (-2.5%)

Embora os resultados do RFQBM sejam baixos P15 e P20, para muitos usuários de sistemas de RI os resultados P5 e P10 são mais interessantes que P15 e P20.

Consideradas, agora, as medidas MAP (Tabela 18) e *R-Precision* (Tabela 19), nota-se que o ganho de efetividade do sistema RFQBM ainda permanece, quando se considera a medida MAP, porém, ao se considerar a medida *R-Precision*, ele é inferior ao *baseline*. Nenhuma das diferenças dos valores de *R-Precision* foi apontada como significativa. Essas medidas, assim como as anteriores, mostram novamente que o RFQBM apresenta a maior proporção documentos relevantes dentre os seus primeiros recuperados (MAP), porém ao considerar um número superior de documentos recuperados, *R-Precision*, sua precisão diminuiu.

Tabela 18 – MAP para as diversas formas de RPR

Sistemas	MAP
RFQBM	**0.4371 (+4.0%)**
RFFullDoc	0.4273 (+1.7%)
RDoc	0.4203
RFGenS	0.4119 (-2.0%)
RFQBS	0.4029 (-4.1%)

Tabela 19 – R-Precision para as diversas formas de RPR

Sistemas	R-Precision
RFFullDoc	0.3949 (+0.6%)
RDoc	0.3923
RFQBM	**0.3826 (-2.4%)**
RFQBS	0.3692 (-5.9%)
Media CLEF 2004	0.3428 (-13%)
RFGenS	0.3278 (-16.4%)

O gráfico da Figura 24 mostra as precisões obtidas com os sistemas RFQBM, RDoc e RFFullDoc. Optou-se por colocar as precisões do RDoc e do RFFullDoc pelo fato do RDoc efetuar a recuperação padrão e o RFFullDoc, a RPR padrão. O gráfico mostra que os valores são muito próximos; porém, quando o grau de revocação aumenta, o RFQBM consegue resultados levemente superiores que o *baseline*.

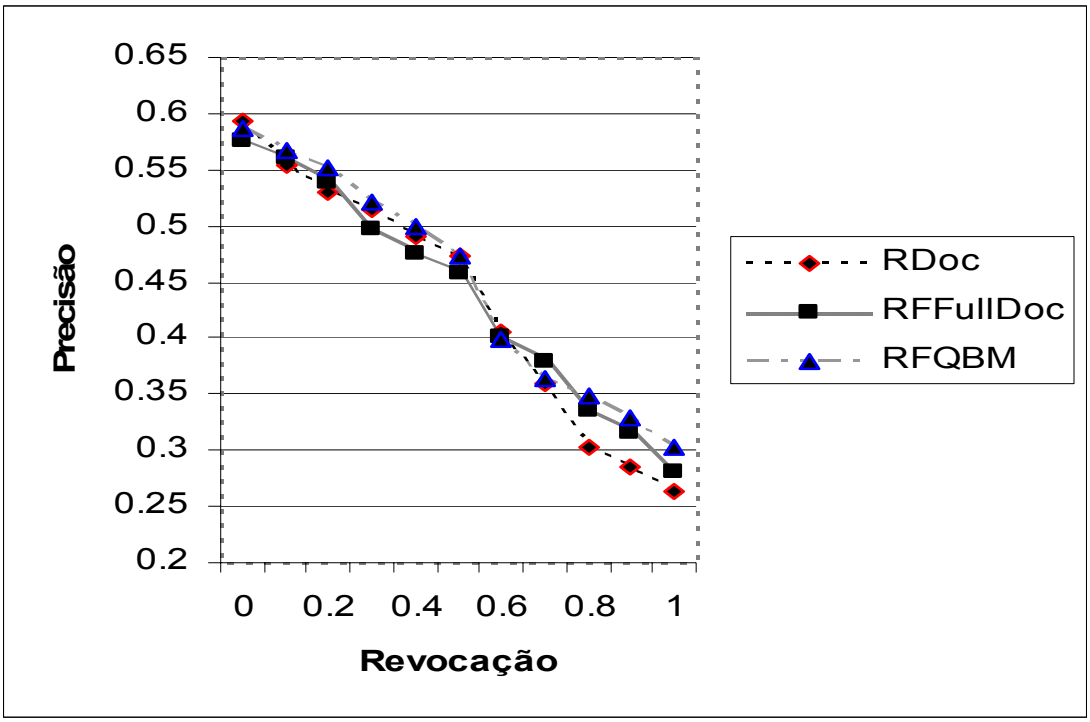


Figura 24 – Precisão para os 11 graus de revocação (interpolada)

Um dos fatores que contribuiu para o baixo incremento da efetividade dos sistemas usando RPR foi que na primeira busca havia poucos documentos relevantes (em média, menos de dois) dentre os cinco pseudo-relevantes documentos. Segundo Buckley e Voorhees (2000), o comportamento dos algoritmos de RPR pode variar de acordo com o número de documentos relevantes por tópico. Tais sistemas podem trabalhar bem para tópicos que levam a muitos documentos relevantes na coleção, mas podem prejudicar a recuperação para tópicos com poucos documentos relevantes correspondentes. Outro fator é que, para 15 tópicos, o RDoc recuperou todos os documentos relevantes nas suas primeiras posições, ou seja, não havia o que melhorar, pois a recuperação inicial já foi perfeita.

Nessa avaliação, houve também casos em que nenhum documento relevante dentre os cinco primeiros documentos recuperados foi encontrado na primeira busca, como para o tópico 201 (Figura 25). Para esse tópico, o sistema RFQBM reformulou a consulta adicionando os seguintes radicais: violent, terc, adquir e destin (radicais das palavras: violentar, terceira, adquiriu e destinar, respectivamente). Claramente esses termos não têm

qualquer relação com o tópico e, portanto, não poderiam contribuir para a melhoria dos resultados. Um conjunto de três dos documentos (irrelevantes) utilizados na reformulação da consulta é mostrado no Anexo 2.

<top>
<num> C201 </num>
<PT-title> Fogos domésticos </PT-title>
<PT-desc> Quais são as principais causas de fogos no lar? </PT-desc>
<PT-narr> Documentos relevantes devem mencionar pelo menos uma causa (provável) de fogos em residencias privadas em geral ou referências a casos específicos.
</PT-narr>
</top>

Figura 25 – Tópico 201 da coleção CHAVE 2004

Assim como ocorreu com o tópico 201, para outros tópicos os sistemas com RPR usaram alguns termos ruins, extraídos de documentos irrelevantes, devido ao problema de que havia muitos documentos irrelevantes no conjunto de pseudo-relevantes.

Um caso singular ocorreu na recuperação para um dos tópicos (215): embora não houvesse nenhum documento relevante nos primeiros cinco documentos recuperados, o sistema RFQBM conseguiu recuperar um documento relevante (o único de toda coleção) na terceira posição da sua lista de documentos recuperados.

Para alguns tópicos com mais de três documentos relevantes dentre os cinco primeiros recuperados, verificou-se que os sistemas selecionaram termos relevantes para a reformulação da consulta inicial. Por exemplo, para o tópico 202 (sobre a prisão de Nick Leeson, Figura 14) os termos relevantes como: baring, singap (radical de Singapura), frankfurt, inglaterr (radical Inglaterra), libr (radical de libra) e broth (radical de brother) foram selecionados pelo RFQBM. Esses termos são fortemente relacionados com o tópico já que Nick Leeson é acusado de falsificação de documentos e fraudes durante sua administração no Bank Baring, banco britânico em Singapura e foi preso em Frankfurt.

Como pode ser notado nesse exemplo, nomes próprios e palavras em língua inglesa tiveram seus radicais extraídos. O fato é que nem os sistemas de recuperação desenvolvidos neste trabalho nem o *stemmer* utilizado por eles identificam nomes próprios ou palavras em outras línguas que não seja a portuguesa. Dessa forma, todas as palavras utilizadas para indexar os documentos e representar as consultas são processadas pelo *stemmer* sem qualquer verificação.

6.2.4 Comparações com os sistemas participantes do CLEF 2004

Pelo fato de a maioria dos usuários de sistemas de RI relutarem em buscar mais do que os dez primeiros documentos com maior similaridade com a consulta (JANSEN et al., 2000), P5 e P10 foram as medidas escolhidas para comparar os sistemas propostos com os sistemas inscritos na tarefa *ad hoc* monolíngüe portuguesa do CLEF 2004[19].

Dos 23 sistemas do CLEF, 16 utilizam os campos *title e description* dos tópicos para formar as consultas a partir das quais os documentos serão recuperados. Como se mencionou antes, os sistemas desenvolvidos neste trabalho também utilizam somente esses campos na avaliação.

As Tabelas 20 e 21 apresentam o desempenho dos sistemas nas medidas P5 e P10, sendo a primeira coluna indicativa de sua classificação. As linhas sombreadas indicam os 7 sistemas construídos neste trabalho, sendo que REext80 e REext60 são duas versões do sistema REext.

[19] Todos os resultados desse fórum, assim como informações sobre os sistemas, podem ser obtidos em http://www.clef-campaign.org, último acesso 13/06/2006.

Tabela 20 – Precisão P5 para a tarefa *ad hoc* do CLEF

Ordem	Sistema	P5
1	UniNEpt2	0.4565
2	UniNEpt1	0.4522
3	UniNEpt3	0.4478
4	humPT04tde	0.4304
5	aplmoptc	0.4130
6	humPT04td	0.4087
7	UAmsC04PoPoAll	0.4043
8	aplmoptb	0.3957
9	tlrpt2	0.3913
10	tlrpt1	0.3870
11	humPT04t	0.3783
12	IRn−MP−Pexp	0.3739
13	**RFQBM**	**0.3720**
14	IRn−MP−Dexp	0.3696
15	aplmopta	0.3696
16	RFQBS	0.3600
17	UAmsC04PoPo4GiSb	0.3565
18	UAmsC04PoPo4GiWd	0.3565
19	RDoc	0.3560
20	RFFullDoc	0.3551
21	RFGenS	0.3480
22	IRn−MP−nexp	0.3391
23	XLDBTumba04	0.2957
24	RExt60	0.2840
25	RDocExt	0.2840
26	XLDBTumba02	0.2783
27	XLDBTumba01	0.2696
28	XLDBTumba05	0.2565
29	RExt80	0.2400
30	Alentejo 1	0.2000
31	Alentejo 2	0.1348

Tabela 21 – Precisão P10 para a tarefa *ad hoc* do CLEF

Ordem	Sistema	P10
1	UniNEpt2	0.3522
2	UniNEpt3	0.3522
3	UniNEpt1	0.3457
4	humPT04tde	0.3326
5	humPT04td	0.3196
6	UAmsC04PoPoAll	0.3174
7	aplmoptc	0.3174
8	aplmoptb	0.3065
9	aplmopta	0.3000
10	UAmsC04PoPo4GiWd	0.2957
11	**RFQBM**	**0.2940**
12	humPT04t	0.2935
13	tlrpt2	0.2935
14	tlrpt1	0.2913
15	RFGenS	0.2880
16	IRn−MP−Pexp	0.2870
17	IRn−MP−Dexp	0.2870
18	RFQBS	0.2840
19	UAmsC04PoPo4GiSb	0.2804
20	RFFullDoc	0.2755
21	RDoc	0.2660
22	IRn−MP−nexp	0.2630
23	RExt60	0.2280
24	RDocExt	0.2160
25	XLDBTumba04	0.2087
26	RExt80	0.2040
27	XLDBTumba02	0.2022
28	XLDBTumba05	0.1870
29	Alentejo 1	0.1652
30	XLDBTumba01	0.1587
31	Alentejo 2	0.1130

Muito embora o RFQBM tenha ficado em primeiro lugar quando comparado somente com os sistemas construídos neste trabalho, ele ocupa uma posição mediana quando a comparação é feita com os sistemas do CLEF. O desempenho inferior aos três melhores sistemas do CLEF pode ser explicado se forem consideradas as seguintes características:

i) uso de recursos de pré-processamento;

ii) método de ponderação de termos;

iii) as próprias estratégias de recuperação.

Assim como os sistemas desenvolvidos neste trabalho, os recursos de pré-processamento utilizados nos três melhores sistemas do CLEF são remoção de *stopwords* e *stemming*. No entanto, a lista de *stopwords* utilizada nos nossos sistemas é composta de conjunções, pronomes e verbos auxiliares, enquanto a dos três melhores sistemas do CLEF é formada por esse tipo de palavras juntamente com um conjunto de 200 palavras mais freqüentes na coleção (por exemplo, a palavra 'público' que é o jornal em que os documentos foram publicados). A remoção dessas 200 palavras pode ter causado uma melhora no desempenho dos três sistemas, pois as palavras mais freqüentes, assim como as demais *stopowords*, são irrelevantes para descrever o conteúdo dos documentos indexados. Em relação ao *stemming*, o utilizado pelos três melhores sistemas do CLEF agrupa na mesma classe de variantes somente plurais e formas de gerúndio, enquanto o utilizado pelos sistemas construídos aqui considera, além dessas variantes, verbos e advérbios. Isso faz com que o *stemming* utilizado pelos sistemas desenvolvidos neste mestrado tenha um número maior de palavras agrupadas na mesma classe. Essa diferença aumenta a probabilidade de *overstemming* (vide Capítulo 3), podendo conseqüentemente prejudicar o processo de recuperação de documentos, devido à baixa precisão que esse tipo de erro pode introduzir.

O método de ponderação utilizado nos três melhores sistemas do CLEF, assim como os desenvolvidos neste trabalho, é baseado na freqüência dos termos nos documentos. No entanto, para normalizar essa freqüência, eles utilizam o tamanho do documento, enquanto o método de ponderação utilizado aqui somente usa o logaritmo. A normalização dos sistemas deles remove a vantagem que documentos longos têm sobre documentos curtos, vantagem causada pelo fato de eles apresentarem um número maior de repetição dos termos, cuja freqüência tende a ser mais alta. Por outro lado, o método de ponderação utilizado aqui reduz o efeito das grandes variações entre as freqüências dos termos de um documento, fazendo

com que a vantagem de documentos longos sobre curtos seja minimizada, mas não removida. A normalização da freqüência dos três sistemas é baseada na seguinte equação:

$$tfn_{ij} = tf_{ij} * \log_2\left(1 + \frac{avgl}{l_i}\right) \quad (25)$$

em que tfn_{ij} é a freqüência normalizada do termo i no documento j, avgl é o tamanho médio de um documento na coleção e l_i é o tamanho do documento i.

Da mesma forma que os sistemas RFQBM, RFQBS e RFGenS, os três melhores sistemas do CLEF utilizaram técnicas de RPR. No entanto, estes utilizam duas estratégias de recuperação e cada uma apresenta uma lista de documentos recuperados. As duas listas são, então, combinadas para formar uma única que será apresentada como saída do sistema. O sistema UniNEPt1 utiliza uma estratégia que faz RPR considerando os 5 primeiros documentos recuperados e extrai 15 termos; a outra faz RPR considerando os 10 primeiros documentos recuperados e extrai 10 termos. O sistema UniNEPt2 utiliza uma estratégia que faz RPR considerando os 5 primeiros documentos recuperados e extrai 30 termos; a outra faz RPR considerando os 10 primeiros documentos recuperados e extrai 15 termos. O sistema UniNEPt3 utiliza uma estratégia que faz RPR considerando os 10 primeiros documentos recuperados e extrai 20 termos; a outra faz RPR considerando os 10 primeiros documentos recuperados e extrai 50 termos. A ponderação dos termos utilizados na RPR é feita de forma similar a Rocchio (1971). O fato de combinar duas estratégias pode ter causado a significativa melhora no desempenho dos sistemas, pois considerando que cada estratégia pode apresentar diferentes documentos relevantes com grande similaridade com a consulta, a combinação das duas estratégias resultaria numa lista com um maior número de documentos relevantes.

Como mostram as tabelas 20 e 21, outros sistemas foram superiores aos sistemas aqui desenvolvidos, mas pareceu-nos interessante compará-los somente com os três primeiros, pois

estes são os que apresentam maiores diferenças de resultados entre si, mostrando que suas características realmente causam um maior impacto na recuperação.

A maioria dos sistemas desenvolvidos neste trabalho apresentou um desempenho melhor que os sistemas Alentejo 1, Alentejo 2, XLDBTumba01, XLDBTumba02 e XLDBTumba05, sendo os três últimos desenvolvidos por um grupo da Faculdade de Ciências da Universidade de Lisboa (FCUL). Diante desse resultado, fez-se uma breve análise dos sistemas desenvolvidos pela FCUL. Embora fosse interessante analisar as características dos sistemas Alentejo 1 e 2, isso não foi feito, pois os dados sobre eles não foram disponibilizados no site do CLEF.

Os três sistemas da FCUL, por sua vez, são baseados na ferramenta de busca Web chamada Tumba[20]. Eles não usam *stemmer* e nem um processo de RPR. Todas as suas consultas são geradas manualmente, baseadas nas informações apresentadas pelos tópicos da coleção. Essas são as diferenças em relação aos sistemas desenvolvidos neste trabalho, pois todos estes usam *stemmer* e suas consultas são construídas automaticamente; além disso, os que apresentam melhores resultados (RFQBM e RFQBS) incorporam a RPR.

O XLDBTumba01 apresenta documentos que foram recuperados para um conjunto de consultas manuais construídas para cada tópico e depois filtrados por dois estudantes de doutorado. Os documentos considerados relevantes por eles foram, então, submetidos ao CLEF 2004. O grupo de FCUL acredita que os resultados ruins são devidos a uma interpretação ruim dos tópicos, resultando em uma construção de consultas e julgamento de relevâncias impróprias.

O XLDBTumba02 recuperou documentos que possuem um *matching* exato com uma consulta manual para cada tópico. Já o sistema XLDBTumba05 usa a distância mínima entre os pares de termos da consulta manual no documento. Dessa forma, documentos que

[20] http://tumba.pt, último acesso em 13/06/2006.

apresentam os termos da consulta com maior proximidade são melhor ranqueados. A diferença entre esses métodos de cálculo de similaridade com o empregado pelos sistemas desenvolvidos neste trabalho pode ser uma das causas das diferenças de desempenho apresentadas.

7 Considerações Finais

Esta dissertação apresentou a verificação do uso de extratos produzidos pelo GistSumm em duas etapas da RI: na indexação e na Realimentação de Pseudo-relevantes (RPR). A motivação para este estudo foi o grau de utilidade dos extratos genéricos do GistSumm obtido na DUC 2003. O objetivo para o seu uso na RPR foi verificar se seus extratos poderiam selecionar dos documentos os termos que possibilitariam melhorar a efetividade da recuperação em comparação com a recuperação sem RPR. Já o objetivo para o seu uso na indexação foi verificar se seus extratos poderiam representar os documentos de forma a proporcionar uma recuperação tão efetiva como se os próprios estivessem sendo utilizados.

A proposta deste trabalho de explorar a contribuição da Sumarização Automática para a RI com o uso de extratos, tanto na indexação quanto na Realimentação de Pseudo-relevantes, resultou na construção e avaliação de cinco sistemas que utilizam o GistSumm para gerar extratos mono e multi-documentos.

Os sistemas que usam extratos na indexação apresentaram uma efetividade inferior ao *baseline*, que utiliza documentos na indexação, mostrando que a utilização de extratos do GistSumm não produziu bons resultados. Os experimentos de Sakai e Sparck-Jones (2001) mostraram que o uso de extratos construídos usando a medida tf-idf ou de extratos *lead* , isto é compostos pelas primeiras sentenças de um documento, também apresentou um desempenho inferior à indexação usando documentos completos. Aqueles extratos só serviram para melhorar a busca por documentos altamente relevantes. Neste trabalho só foi experimentada a busca por documentos julgados com relevância binária (relevante ou irrelevante). Assim, não foi possível testar se os extratos do GistSumm são úteis para a recuperação de documentos altamente relevantes.

Analisando alguns dos documentos utilizados nos experimentos, constatou-se que havia muitos considerados relevantes cuja informação relevante não era o seu assunto principal e outros com várias matérias jornalísticas de diferentes temas. Os extratos desses documentos gerados pelo GistSumm, muitas vezes, não continham as informações relevantes para os tópicos, já que, no primeiro caso, as informações não pertenciam a idéia principal do documento, e no segundo caso, as informações pertenciam a uma das matérias que não foram consideradas como assunto principal pelo GistSumm. Dessa forma, os índices gerados a partir dos extratos não possibilitaram a recuperação de documentos relevantes, fazendo com que a performance dos sistemas que usam extratos na indexação fosse inferior ao que usa documentos. Isso indica que os extratos do GistSumm são pouco úteis para indexação de documentos quando esses possuem múltiplos tópicos e quando a informação relevante aos tópicos de busca é superficial.

Nos experimentos com RPR, três tipos de extratos foram utilizados: genéricos mono-documento, específicos mono-documento e específicos multi-documentos. O uso dos extratos específicos multi-documentos (RFQBM) na RPR não foi encontrado em nenhuma publicação. Os resultados dos sistemas que utilizam esses tipos de extratos são comparados com um sistema sem RPR (*baseline*) e com outro que utiliza documentos na RPR (RFFullDoc). Em geral, todos esses resultados são muito próximos. A provável justificativa é o número pequeno de documentos relevantes por tópico da coleção utilizada. Para 22 tópicos dos 50 da coleção havia no máximo três documentos relevantes. Além disso, para 15 tópicos, a recuperação inicial, sem RPR, já havia atingindo o melhor resultado possível, recuperando todos os documentos relevantes nas primeiras posições.

Quando números pequenos de primeiros documentos recuperados (5 e 10) são considerados, a precisão do RFQBM é maior que as precisões dos outros sistemas, mostrando que seus termos contribuíram para uma melhora da recuperação. Considerando os mesmos

números de documentos recuperados, o RFQBS (que utiliza extratos específicos mono-documentos) tem uma precisão superior ao *baseline* e ao RFFullDoc. Já o RFGen (que usa extratos genéricos mono-documentos) tem o pior resultado para os primeiros cinco documentos recuperados, porém apresenta um melhor desempenho quando são considerados 10 documentos. Em resumo, quando poucos documentos são considerados, os sistemas que usam extratos específicos, em especial os multi-documentos, apresentam maior proporção de documentos relevantes recuperados.

Para um número maior de primeiros documentos recuperados (15 e 20), o RFQBM não apresenta diferenças significativas dos demais sistemas. Um fato que pode explicar isso é a seleção de termos muito específicos, pertencentes a um ou dois documentos pseudo-relevantes, que são acrescentados a algumas das consultas reformuladas. Apesar da pouca diferença, todos os sistemas que utilizam extratos na RPR apresentam precisões superiores às do *baseline* e às do RFFullDoc.

Quando é considerado o número de documentos relevantes da coleção (*R-Precision*), o RFQBM é o sistema que apresenta a maior proporção de documentos relevantes recuperados em comparação com os sistemas que utilizam extratos. A medida MAP mostra que o RFQBM foi o sistema que apresentou os documentos relevantes nas melhores posições da lista de documentos recuperados.

Com base nos resultados expostos aqui, os extratos multi-documentos específicos, para a RPR, indicam um bom potencial de melhora de desempenho da recuperação, apresentando um número maior de documentos relevantes, principalmente dentre os primeiros documentos recuperados. O fato de usuários de sistemas de RI geralmente verificarem um número pequeno de documentos recuperados, indica que o RFQBM pode apresentar maior aceitabilidade que os demais.

Ainda, considerando que o sumarizador extrativo utilizado, o GistSumm, é baseado em um método muito rudimentar de geração de extratos multi-documentos, o uso do RFQBM com um sumarizador mais expressivo, como aqueles que tratam a redundância da informação (por exemplo, Goldstein et al., 2000), pode ainda melhorar o desempenho da RI, especialmente considerando a RPR.

O fato de os testes de significância, teste dos sinais e teste t apresentarem como significativa somente uma diferença entre os resultados dos sistemas com RPR indica que as diferenças de desempenho para cada tópico são pequenas. A justificativa continua sendo o número de documentos relevantes da coleção utilizada; esses resultados são, de certa forma, esperados. Os resultados desses testes indicam que experimentos com coleções com um número maior de documentos relevantes devem ser considerados, a fim de verificar se as diferenças de desempenho dos sistemas são mantidas.

Nas próximas seções serão apresentadas as contribuições deste trabalho, suas limitações e também possíveis desdobramentos deste trabalho.

7.1 Contribuições deste trabalho

Embora os resultados deste trabalho sejam modestos, ele possibilita o desenvolvimento de várias etapas importantes para a recuperação de documentos, fornecendo recursos para a exploração de uma metodologia importante atualmente: a realimentação de pseudo-relevantes. Além disso, a construção e avaliação dos sistemas demandaram a disponibilização de processos e dados complementares que poderão ser usados também em outros projetos. Os principais índices de produtividade deste trabalho indicam esse potencial:

1. disponibilização de sete sistemas de Recuperação de Informação em Java, que implementam diversos métodos de indexação e realimentação de pseudo-relevantes.

2. indicação de um sistema potencialmente útil, baseado em extratos específicos multi-documentos, para a recuperação de documentos em português.
3. disponibilização de corpora de extratos dos documentos da coleção CHAVE 2004, produzidos pelo GistSumm, incluindo:
    - um corpus de 55.070 extratos mono-documentos com taxa de compressão de 80%;
    - um corpus de 55.070 extratos mono-documentos com taxa de compressão de 60%;
4. disponibilização do mecanismo de construção de arquivos de índices invertido, usando diversas técnicas, para investigações futuras.
5. implementação do *stemmer* de Orengo e Huyck (2001) em Java, para o português, antes só disponível em C^{++}.

7.2 Limitações deste trabalho

Os experimentos realizados neste mestrado são modestos e limitados a um contexto bastante restrito, quer em relação ao corpus de teste (textos jornalísticos, coleção relativamente pequena, se considerada a tarefa de RI), quer em relação ao tipo de tarefa considerado na avaliação. No entanto, vale ressaltar que ambas as limitações são comuns e necessárias para a avaliação de desempenho de sistemas de RI, tanto que são características do próprio CLEF. O fato de adotar-se esse contexto para avaliar os sistemas propostos licencia e justifica os resultados, embora não exclua a necessidade de testes mais robustos.

Testes mais abrangentes também são necessários, principalmente se for considerado que, em um ambiente real de RI, como o da Web, os textos encontram-se formatados e esta não foi uma característica aproveitada neste trabalho. Documentos em HTML, por exemplo,

podem apresentar informações importantes sobre a estrutura e o conteúdo do texto, para a sumarização automática e, conseqüentemente, para a melhoria dos índices de recuperação. Certamente, o desempenho dos sistemas com coleções de documentos diferentes ou com textos formatados pode ser muito diverso do apresentado neste trabalho.

Outras técnicas mais elaboradas de processamento textual, como a etiquetação morfossintática ou métodos de resolução de anáforas (como o de Edens et al., 2003), que reconhecidamente podem aumentar a efetividade da recuperação melhorando a seleção e ponderação de termos indexadores, não foram consideradas aqui porque se privilegiou maior independência, em relação ao processamento de língua natural. Pelo mesmo motivo, mecanismos de identificação de termos compostos, que, em geral, carregam mais informação que termos simples, não foram considerados. Além disso, outras medidas, quer de similaridade, quer de ponderação, também poderiam ser exploradas.

Uma grande limitação dos sistemas propostos está na sua forma de indexação: não houve um investimento efetivo nessa etapa, embora tenham sido utilizados modelos licenciados para a criação do arquivo de índices invertido. Eles são, inclusive, pouco eficientes com relação a algumas ferramentas de buscas reais. Por exemplo, a Terrier, desenvolvida na Universidade de Glasgow, foi implementada com um bom algoritmo de compressão de índices que possibilita o carregamento parcial do arquivo em memória, agilizando o processo de busca. Os sistemas aqui descritos são dispendiosos, quer em tempo de processamento, quer em espaço de memória.

No que diz respeito ao cálculo da similaridade entre documentos e consultas, a proposta deste trabalho também é simplificada: outras medidas que não a simples soma das freqüências dos termos no documento e no extrato, como no RDocExt, poderiam ter sido utilizadas. Além disso, a coincidência de valores de similaridade exigiria uma técnica de desempate, que não foi considerada, mas sabidamente poderia melhorar o desempenho de

alguns sistemas, como o RExt. Este problema se torna ainda mais grave quando a coincidência ocorre entre documentos relevantes e irrelevantes. Uma forma de desempate simples poderia ser o próprio tamanho dos documentos com mesmos graus de similaridade, de fácil implementação e, talvez, com resultados finais mais promissores.

Por fim, os sistemas implementados possibilitam a apresentação dos extratos dos documentos recuperados, porém, nenhum teste da funcionalidade dos extratos em si foi realizado, ou seja, não foi verificado se os extratos do GistSumm efetivamente facilitam o julgamento de relevância dos documentos correspondentes, tópico de grande importância na associação das duas grandes áreas em foco neste trabalho: a Sumarização Automática e a Recuperação de Informação. Uma forma de fazer isso seria submeter extratos e documentos ao julgamento de um comitê de usuários (p.ex., como em TOMBROS; SANDERSON, 1998 ou AMITAY, 2001). Esta forma, aliás, seria pouco custosa, já que todo o ambiente para esse tipo de avaliação já foi disponibilizado no NILC (PEDREIRA-SILVA, 2006).

7.3 Trabalhos futuros

Os sistemas e avaliações apresentados aqui podem ser estendidos e aprimorados de diversas formas. Para a recuperação usando extratos na indexação, uma tarefa interessante seria verificar se o GistSumm poderia auxiliar na tarefa de recuperar documentos altamente relevantes. Para isso, seria necessária uma coleção de documentos marcados com seus graus de relevância previamente. Esta opção, no entanto, é de difícil execução, por envolver um julgamento de difícil confiabilidade.

Outra opção seria verificar se os extratos podem auxiliar em buscas com consultas limitadas, similar ao sistema SEARCHABLE LEAD (WASSON, 1998), em que o usuário tem a opção de limitar as partes de sua consulta que necessitam estar nas sentenças iniciais do

documento. Este caso seria similar à construção de extratos usando as sentenças *lead.* Os componentes da consulta não delimitados não seriam restritos a qualquer localização no documento. Para a avaliação, seria interessante, então, verificar se os extratos do GistSumm delimitados a segmentos da consulta levariam a um desempenho melhor do que aquele baseado em consultas não limitadas. O problema desta estratégia seria contar com a interatividade de usuários, em vez de simplesmente efetuar avaliações automáticas com as grandes coleções de testes com tópicos e julgamentos já definidos, como o corpus do CLEF 2004 utilizado neste trabalho.

Ainda para a indexação, usando extratos genéricos seria interessante considerar um sumarizador que extraísse segmentos de um documento envolvendo diferentes tópicos. Isso poderia resolver o problema de os extratos nem sempre contribuírem para a recuperação de documentos com múltiplos tópicos. Esta perspectiva é real, pois já está prevista uma nova versão do GistSumm que atenderá esse requisito.

Na realimentação de pseudo-relevantes, existem variações a serem consideradas no uso de extratos como meio de realimentação, tais como: o número de documentos pseudo-relevantes, o número de termos selecionados, a taxa de compressão dos extratos e também a escolha de diversas funções de seleção de termos. Porém, deve-se ressaltar que alguns desses parâmetros, como a taxa de compressão e número de termos selecionados, são dependentes entre si. Não adianta permitir um número grande de termos se a limitação de tamanho do extrato não permitir sua inclusão integral.

A própria realimentação de relevantes (RR), isto é, a realimentação real a partir do julgamento de documentos exibidos aos usuários (e, portanto, com sua interação em tempo real com o sistema) também constitui outro aspecto importante a investigar, potencializado por este trabalho. Uma vez possibilitada, haveria ainda outra forma de avaliar os sistemas propostos aqui: considerando os resultados automáticos obtidos na RR como corpus de

referência e, então, comparando os demais conjuntos de resultados com este, para cálculo de precisão ou revocação, por exemplo.

Considerando o potencial do sistema RFQBM (RPR com extratos específicos multi-documentos), uma extensão interessante deste mestrado será incorporar ao sistema um sumarizador multi-documentos mais sofisticado, que atenda os requisitos das metodologias de ponta na área de Sumarização Automática Multi-documentos. Trabalhos nessa linha já estão em andamento no NILC e poderão ser considerados também no contexto deste mestrado.

Considerando o estado atual dos sistemas apresentados, outras tarefas de avaliação podem ser consideradas, independentemente de quaisquer melhorias. Por exemplo, adotando coleções de documentos mais significativas para a recuperação de documentos em português, como a do CLEF 2006. Diferentemente da coleção CLEF 2004, esta é formada por mais de 209 mil documentos escritos tanto em português europeu (artigos do jornal PÚBLICO) quanto em português brasileiro (artigos do jornal Folha de São Paulo). Aliás, já se encontram em fase de submissão ao CLEF atual os resultados de alguns sistemas aqui apresentados, incluindo o RDoc e RFQBM[21]. Como conclusão dessa etapa, será possível redirecionar as pesquisas iniciadas com esse mestrado e delinear mais claramente quais, dentre as linhas de progresso apontadas, serão de fato mais promissoras.

Por fim, como o GistSumm também incorpora um módulo de sumarização de textos em inglês, seria interessante verificar o desempenho dos sistemas para coleções de documentos nessa língua. Os recursos de pré-processamento não seriam problema, já que todos eles estão amplamente disponíveis. Tampouco os modelos de manipulação da informação, quer na sumarização, quer na própria construção dos graus de relevância para a reformulação da consulta e determinação de documentos relevantes de uma coleção, seriam

[21] Cabe ressaltar que até 2005 nenhum grupo brasileiro havia participado da tarefa de recuperação monolíngüe de documentos em português do CLEF. Aliás, a língua portuguesa foi incorporada a esse concurso somente em 2004, com a coleção usada também neste mestrado.

problemáticos, já que são independentes de língua natural. Por último, a utilização de sistemas baseados em outro modelo que não o vetorial poderia também ser considerada.

Referências bibliográficas

ADAFRE, S.F. et al. The University of Amsterdam at CLEF 2004. In: CLEF 2004 WORKSHOP, 5., 2004, Bath. *Proceedings*… London: Springer-Verlag, 2004. p. 91–98.

AMITAY, E. What lays in the layout: using anchor-paragraph arrangements to extract descriptions of Web documents. 2001. 147 p. Tese (Doutorado) - Division of Information and Communication Sciences, Macquarie University, Sydney, 2001.

BRANDOW, R.; MITZE, K.; RAU, L.F. Automatic condensation of electronic publications by sentence selection, *Information Processing and Management*, v.31, n. 5, p. 675–685, 1995.

BAEZA-YATES, R.A.; RIBEIRO-NETO, B. *Modern information retrieval*. Boston: Addison-Wesley, 1999. 513p.

BRASCHLER, M.; PETERS, C. Cross-language evaluation forum: objectives, results, achievements, *Information Retrieval*, v.7, n. 1-2., p. 7–31, 2004.

BELEW, R.K. *Finding out about*: a cognitive perspective on search engine technology and the WWW. New York: Cambridge University Press, 2000. 384p.

BUCKLEY, C.; VOORHEES, E.M. Evaluating evaluation measure stability. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23., 2000, Athens. *Proceedings*… New York: ACM Press, 2000. p. 33–40.

CALDAS JUNIOR, J.; IMAMURA, C.Y.M.; REZENDE, S.O. Avaliação de um algoritmo de stemming para a língua portuguesa. In: CONGRESS OF LOGIC APPLIED TO TECHNOLOGY, 2., 2001, São Paulo. *Proceedings*...São Paulo: Faculdade Senac de Ciências Exatas e Tecnologia, 2001. p. 267–274.

CARPINETO, C.; ROMANO, G. Towards more effective techniques for automatic query expansion. In: EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 3., 1999, Vienna. *Proceedings...* London: Springer-Verlag, 1999. p.126–141.

CHAVES, M.S. Um estudo e apreciação sobre algoritmos de stemming. In: JORNADAS IBEROAMERICANAS DE INFORMÁTICA, 9., 2003, Cartagena de Índias. Disponível em:

<http://xldb.di.fc.ul.pt/~mchaves/pg_portugues/public/stemming.pdf>, Acesso em: maio 2006.

CORMACK, G.V.; PALMER, C.R.; CLARKE, C.L.A. Efficient construction of large test collections. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., 1998, Melbourne. *Proceedings*...New York: ACM Press, 1998. p. 282–289.

EDENS, R.J.et al. An investigation of broad coverage automatic pronoun resolution for information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 26., 2003, Toronto. *Proceedings*... New York: ACM Press, 2003. p. 381–382.

ELLIOTT, R.W.; CASHMAN, L.E. An experimental comparison of relevance-feedback techniques. In: ANNUAL INTERNATIONAL ACM CONFERENCE, 1973, Atlanta. *Proceedings*... New York: ACM Press, 1973. p. 256–261.

ELMASRI, R.; NAVATHE, S.B. *Fundamentals of database systems*. 3. ed. Menlo Park: Addison-Wesley, 2000. 908 p.

FAGAN, J. Automatic phrase indexing for document retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 10., 1987, New Orleans. *Proceedings*... New York: ACM Press, 1987. p. 91–101.

FULLER, M.; ZOBEL, J. Conflation-based comparison of stemming algorithms. In: AUSTRALIAN DOCUMENT COMPUTING SYMPOSIUM, 3., Sydney, 1998. p. 8-13.

GOLDSTEIN, J. et al. Creating and evaluating multi-document sentence extract summaries. In: CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 9., 2000, Virginia. *Proceedings*...New York: ACM Press, 2002. p. 165–172.

HARMAN, D., Relevance feedback revisited. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 15., 1992, Copenhagen. *Proceedings*...New York: ACM Press, 1992. p. 1–10.

HARMAN, D. Overview of the third Text Retrieval Conference. In: TEXT RETRIEVAL CONFERENCE, 3., 1994, Gaithersburg. *Proceedings*... Gaithersburg: NIST Special Publication, 1995. p. 1–20.

HARMAN, D. Overview of the fourth Text Retrieval Conference. In: TEXT RETRIEVAL CONFERENCE, 4., 1995, Gaithersburg. *Proceedings*... Gaithersburg: NIST Special Publication, 1996. p. 1–23.

HULL, D. Using statistical testing in the evaluation of retrieval experiments. In: ANNUAL ACM INTERNATIONAL SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 16., 1993, Pittsburgh. *Proceedings*... New York: ACM Press, 1993. p. 329–338.

JANSEN, B.J.; SPINK, A.; SARACEVIC, T. Real life, real users, and real needs: a study and analysis of user queries on the web, *Information Processing and Management*, v. 36, n. 2, p. 207–227, 2000.

KENT, A. et al. Machine literature searching VIII: operational criteria for designing information retrieval systems, *American Documentation*, v. 6, n. 2, p. 93–101, 1955.

LAM-ADESINA, A.M.; JONES, G.J.F. Applying summarization techniques for term selection in relevance feedback. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 24., 2001, New Orleans. *Proceedings*... New York: ACM Press, 2001. pp. 1–9.

LEE, J.H. Combining multiple evidence from different properties of weighting schemes. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 18., 1995, Seattle. *Proceedings*...New York: ACM Press, 1995. p.180–188.

LLOPIS F. et al. IR-n r2: using normalized passages. In: CLEF 2004 WORKSHOP, 5., 2004, Bath. *Proceedings*… London: Springer-Verlag, 2004. p. 65–72.

LOVINS, J. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, v. 11, n. 1, p. 22–31, 1968.

LUHN, H.P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, v. 2, n. 2 , p.159–165, 1958.

MANDALA, R.; TOKUNAGA, T.; TANAKA, H. Combining multiple evidence from different types of thesaurus for query expansion. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 22., 1999, Berkeley. *Proceedings*... New York: ACM Press, 1999. p. 191–197.

MANI, I. *Automatic summarization*. Philadelphia: John Benjamins, 2001. 286 p.

McKEOWN, K.R. et al. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE, 2002, San Diego. Disponível em: <http://newsblaster.cs.columbia.edu/papers/hlt-blaster.pdf> Acesso em: maio 2006.

MILLER, G.A. WordNet: a lexical database for English. *Communications of the ACM*, v.38, n. 11, p. 39–41, 1995.

MONTGOMERY, J. et al. Effect of varying number of documents in blind feedback: analysis of the 2003 NRRC RIA workshop "bf_numdocs" experiment suite. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 27., 2004, Sheffield. *Proceedings*...New York: ACM Press, 2004. p. 476–477.

NAVARRO, G. Indexing and searching. In: BAEZA-YATES, R.A.; RIBEIRO-NETO, B. *Modern information retrieval*. Boston: Addison-Wesley, 1999. p.191–228.

ORENGO, V.M.; HUYCK, C. A stemming algorithm for Portuguese language. In: SYMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL, 8., 2001, Laguna de San Rafael. *Proceedings*… New York: IEEE Computer Society Publications, 2001. p. 183–193.

PARDO, T.A.S. *GistSumm:* um sumarizador automático baseado na idéia principal de textos. São Carlos-SP: USP/ICMC, 2002. Série de Relatórios do NILC. NILC-TR-02-13. 25 p.

PARDO, T.A.S.; RINO, L.H.M.; NUNES, M.G.V. GistSumm: a summariztion tool base on a new extractive method. In: WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 6., 2003, Faro. *Proceedings*…Germany: Springer-Verlag, 2003. p. 210–218.

PARDO, T.A.S. *GistSumm*: GIST SUMMarizer: extensões e nova funcionalidades. São Carlos-SP: USP/ICMC, 2005. Série Relatórios do NILC. 8 p.

PEDREIRA-SILVA, P. ExtraWeb: um sumarizador de documentos Web baseado em etiquetas HTML e ontologia. Dissertação de Mestrado, em fase de conclusão. Departamento de Computação, UFSCAR. São Carlos, SP, 2006.

PORTER, M.F. An algorithm for suffix stripping. *Program*, v. 14, n. 3, p. 130–137, 1980.

RAGHAVAN, V.V.; JUNG, G.S.; BOLLMANN, P. A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Transactions on Information Systems*, v. 7, n. 3, p.205–229, 1989.

ROBERTSON, S.E.; SPARCK-JONES, K. Relevance weighting on search terms. *Journal of American Society for Information Sciences*, v. 27, n. 3, p.129–146, 1976.

ROBERTSON, S.E., On term selection for query expansion. *Journal of Documentation*, v. 46, n. 4, p. 359–364, 1990.

ROCCHIO JR., J. J.. Relevance feedback in information retrieval. In: SALTON, G. *The SMART retrieval system:* experiments in automatic document processing. New Jersey: Prentice Hall, 1971. p. 313–336.

RUTHVEN, I. Re-examining the potential effectiveness of interactive query expansion, In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 26., 2003, Toronto. *Proceedings*...New York: ACM Press, 2003. p. 213–220.

SAKAI, T.; SPARCK-JONES. Generic summaries for indexing in information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 24., 2001, New Orleans. *Proceedings*... New York: ACM Press, 2001. p. 190–198.

SALTON, G. Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, v. 10, n. 4, p. 440–457, 1963.

SALTON, G. *Automatic information organization and retrieval*. New York: McGraw Hill, 1968. 421 p.

SALTON, G. *The SMART Retrieval System:* experiments in automatic document processing. New Jersey: Prentice Hall, 1971. 548 p.

SALTON, G.; McGILL, M.J., *Introduction to modern information retrieval*. New York: McGraw-Hill, 1983. 448 p.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management ,* v. 24, n. 5, p. 513–523, 1988.

SANDERSON, M.; JOHO, H. Forming test collections with no system pooling. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 27., 2004, Sheffield. *Proceedings*...New York: ACM Press, 2004. p. 33–40.

SANTOS, D.; ROCHA, P. CHAVE: topics and question on the portuguese participation in CLEF. In: CLEF 2004 WORKSHOP, 5., 2004, Bath. *Proceedings*… London: Springer-Verlag, 2004. p. 639–648.

SARACEVIC, T. Evaluation of evaluation in information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 18., 1995, Seatle. *Proceedings*...New York: ACM Press, 1995. p. 138–146.

SCHIFFMAN, B.; NENKOVA, A.; MCKEOWN, K. Experiments in multidocument summarization. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE, 2002, San Diego. Disponível em: <http://newsblaster.cs.columbia.edu/papers/hlt02-dems.pdf>. Acesso em: maio 2006.

SCHOLER, F. et al. Compression of inverted indexes for fast query evaluation. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 25, 2002, Tampere. *Proceedings*...New York: ACM Press, 2002. p. 222–229.

SORMUNEN, E. Liberal relevance criteria of TREC - counting on negligible documents?. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 25., 2002, Tampere. *Proceedings*...New York: ACM Press, 2002. p. 324–330.

SPARCK-JONES, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, v. 28, n. 1 , p. 11–20, 1972.

SPARCK-JONES, K.; Van RIJSBERGEN, C.J. *Report on the need for and provision of an 'ideal' information retrieval test collection*. New York: University Computer Laboratory, 1975. British Library Research and Development Report 5266.

TOMBROS, A.; SANDERSON, M. Advantages of query biased summaries in information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., 1998, Melbourne. *Proceedings*... New York: ACM Press, 1998. p. 2–10.

TRIOLA, M.F. *Introdução à estatística*. Rio de Janeiro: LTC Editora, 1999. 410 p.

Van RIJSBERGEN, C.J. *Information Retrieval*. 2. ed. Newton: Butterworth-Heinemann., 2nd edition, 1979. 208p.

VOORHEES, E.M. Variations in relevance judgments and the measurement of retrieval effectiveness. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., 1998, Melbourne. *Proceedings*... New York: ACM Press, 1998. p. 315–323.

VOORHEES, E. M.; HARMAN, D. Overview of the seventh text retrieval conference. In: TEXT RETRIEVAL CONFERENCE, 7., 1998, Gaithersburg. *Proceedings*... Gaithersburg: NIST Special Publication. p. 1–23, 1998.

VOORHEES, E. M.; HARMAN, D. Overview of the eighth text retrieval conference. In: TEXT RETRIEVAL CONFERENCE, 8., 1999, Gaithersburg. *Proceedings*... Gaithersburg: NIST Special Publication. p. 1–23, 1999.

WASSON, M. Using leading text for news summaries: evaluation results and implications for commercial summarization applications. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 17., 1998, Montreal. *Proceedings*... Morristown: Association for Computational Linguistics, 1998. p. 1364–1368.

WHITE, R.W.; RUTHVEN, I.; JOSE, J.M. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 25., 2002, Tampere. *Proceedings*... New York: ACM Press, 2002. p. 57–64.

XU, J.; CROFT, W.B. Query expansion using local and global document analysis. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 19., 1996, Zurich. *Proceedings*... New York: ACM Press, 1996. p. 4–11.

XU, J.; CROFT, W.B. Corpus-based stemming using co occurrence of word variants. *ACM Transactions on Information Systems*, v. 16, n. 1, p. 61–81, 1998.

Anexo 1 – Tópicos da CHAVE 2004

<top>
<num> C201 </num>
<PT-title> Fogos domésticos </PT-title>
<PT-desc> Quais são as principais causas de fogos no lar? </PT-desc>
<PT-narr> Documentos relevantes devem mencionar pelo menos uma causa (provável) de fogos em residencias privadas em geral ou referências a casos específicos. </PT-narr>
</top>

<top>
<num> C202 </num>
<PT-title> Prisão de Nick Leeson </PT-title>
<PT-desc> Encontrar documentos sobre a prisão de Nick Leeson e as causas que o levaram à cadeia. </PT-desc>
<PT-narr> Documentos relevantes devem relatar as razões da prisão de Nick Leeson e sua subsequente prisão.
</PT-narr>
</top>

<top>
<num> C203 </num>
<PT-title> Guerrilha em Timor Leste </PT-title>
<PT-desc> Encontrar documentos sobre actividade de guerrilha em Timor Leste. </PT-desc>
<PT-narr> Documentos relevantes devem mencionar actividades concretas políticas ou militares das forças de guerrilha leste-timorenses. </PT-narr>
</top>

<top>
<num> C204 </num>
<PT-title> Vítimas de Avalanches </PT-title>
<PT-desc> Encontrar informação sobre o número de mortos em avalanches. </PT-desc>
<PT-narr> Documentos relevantes devem detalhar o número de pessoas que morreram em avalanches; tanto documentos sobre avalanches individuais como estátisticas mais gerais sobre mortes causadas por avalanches são relevantes. </PT-narr>
</top>

<top>
<num> C205 </num>
<PT-title> Ataques suicidas tamil </PT-title>
<PT-desc> Encontrar algumas informação sobre ataques bombistas suicidas dos Tigres Tamil ou acções kamikazes no Sri Lanka. </PT-desc>
<PT-narr> Apenas documentos sobre ataques bombistas suicidas por rebeldes tamil são relevantes; outras formas de ataque não são importantes. </PT-narr>
</top>

<top>
<num> C206 </num>
<PT-title> Cimeira dos G7 em Halifax </PT-title>
<PT-desc> Encontrar documentos sobre as discussões sobre as reformas de instituições financeiras, em particular o Banco Mundial e o FMI, na cimeira dos G7 que se realizou em Halifax em 1995. </PT-desc>
<PT-narr> Documentos relevantes devem detalhar as discussões económicas e mencionar algumas propostas para a reforma das instituições financeiras internacionais. </PT-narr>
</top>

<top>
<num> C207 </num>
<PT-title> Ferimentos provocados por fogo de artifício </PT-title>
<PT-desc> Encontrar documentos relatando mortes ou ferimentos causados por fogos de artifício. </PT-desc>

<PT-narr> Documentos relevantes devem relatar mortes ou ferimentos causados por fogos de artifício. Tanto documentos fornecendo estatísticas de ferimentos como relatórios de acidentes singulares são de interesse. </PT-narr>
</top>

<top>
<num> C208 </num>
<PT-title> "O Mundo de Sofia" </PT-title>
<PT-desc> Encontrar documentos sobre o sucesso editorial do livro "O Mundo de Sofia" de Jostein Gaarder. </PT-desc>
<PT-narr> Documentos relevantes devem descrever o tema de "O Mundo de Sofia", e devem mencionar o seu sucesso de vendas. </PT-narr>
</top>

<top>
<num> C209 </num>
<PT-title> Vencedor da Volta a França </PT-title>
<PT-desc> Quem venceu a Volta a França de 1995? </PT-desc>
<PT-narr> Documentos relevantes devem mencionar o nome do vencedor da Volta a França em 1995. </PT-narr>
</top>

<top>
<num> C210 </num>
<PT-title> Candidatos ao Prémio Nobel da Paz </PT-title>
<PT-desc> Encontrar documentos discutindo os nomes de qualquer dos candidatos ao Prémio Nobel da Paz de 1995. </PT-desc>
<PT-narr> Documentos devem reflectir previsões prévias ao anúncio do Prémio Nobel da Paz relativas a possíveis vencedores. Documentos que apenas mencionem o vencedor não são relevantes. </PT-narr>
</top>

<top>
<num> C211 </num>
<PT-title> Conflito fronteiriço Peru-Equador </PT-title>
<PT-desc> Encontrar documentos sobre a disputa fronteiriça entre o Peru e o Equador. </PT-desc>
<PT-narr> Documentos relevantes devem fornecer pormenores sobre a disputa fronteiriça entre o Peru e o Equador. </PT-narr>
</top>

<top>
<num> C212 </num>
<PT-title> Mulheres desportistas e doping </PT-title>
<PT-desc> Encontrar documentos falando sobre mulheres que foram acusadas de usar substâncias dopantes para melhorar os seus resultados desportivos. </PT-desc>
<PT-narr> Quer documentos discutindo o problema geral de mulheres desportistas usando doping quer documentos mencionando os nomes das mulheres acusadas são relevantes. </PT-narr>
</top>

<top>
<num> C213 </num>
<PT-title> Viagens papais </PT-title>
<PT-desc> Encontrar documentos sobre viagens fora de Itália realizadas pelo Papa João Paulo II em 1995. </PT-desc>
<PT-narr> Documentos relevantes mencionarão qualquer dos países na Europa, África ou Ásia visitados pelo Papa em 1995. Menções a viagens realizadas noutros anos não são relevantes. </PT-narr>
</top>

<top>
<num> C214 </num>
<PT-title> Multi-bilionários </PT-title>

<PT-desc> Encontrar documentos sobre a riqueza de multi-bilionários em qualquer lugar do mundo. </PT-desc>
<PT-narr> Documentos relevantes devem citar o nome de multi-bilionários individuais detalhando o valor da sua fortuna ou devem fornecer valores gerais sobre multi-bilionários no mundo. </PT-narr>
</top>

<top>
<num> C215 </num>
<PT-title> Reeleição do presidente do Peru </PT-title>
<PT-desc> Encontrar documentos sobre a reeleição do presidente peruano em 1995. </PT-desc>
<PT-narr> Documentos relevantes devem mencionar a reeleição, ou seja, o segundo mandato. </PT-narr>
</top>

<top>
<num> C216 </num>
<PT-title> Inalação de cola por jovens </PT-title>
<PT-desc> Documentos sobre a popularidade da inalação de cola por crianças e adolescentes </PT-desc>
<PT-narr> Documentos devem relatar incidentes específicos ou fornecer informação relacionada com o hábito de inalar cola por crianças e adolescentes. Informação sobre legislação destinada a proibir o fornecimento de kits de inalação de cola a crianças também é relevante. A inalação de outras substâncias não é de interesse. </PT-narr>
</top>

<top>
<num> C217 </num>
<PT-title> Sida em África </PT-title>
<PT-desc> Encontrar documentos discutindo o crescimento da sida em África. </PT-desc>
<PT-narr> Houve um aumento explosivo da sida em África. Documentos relevantes discutirão este problema. De particular interesse são documentos mencionando organizações humanitárias lutando contra a sida em África. </PT-narr>
</top>

<top>
<num> C218 </num>
<PT-title> Andreotti e a Mafia </PT-title>
<PT-desc> Encontrar informação sobre acusações judiciais apresentadas contra o ex-primeiro-ministro italiano, Giulio Andreotti, acusando-o de pertencer ou de estar associado à Mafia. </PT-desc>
<PT-narr> Documentos relevantes devem detalhar as acções legais; a mera listagem das acusações sem informação adicional não é suficiente </PT-narr>
</top>

<top>
<num> C219 </num>
<PT-title> Candidatos a Comissário Europeu </PT-title>
<PT-desc> Quais as razões dadas pelos deputados europeus para não aceitar totalmente quatro candidatos à Comissão e qual a nacionalidade desses candidatos? </PT-desc>
<PT-narr> Documentos relevantes devem mencionar as razões das objecções a certos candidatos e as nacionalidades de tais candidatos. Exemplos relativos a um único candidato também são pertinentes. Documentos descrevendo a audição de candidatos também são de interesse. </PT-narr>
</top>

<top>
<num> C220 </num>
<PT-title> Carros europeus na Rússia </PT-title>
<PT-desc> Encontrar documentos relatando a exportações de carros europeus para a Rússia. </PT-desc>
<PT-narr> Documentos relevantes tratam da exportação de carros para a Rússia e devem fazer menção específicas a alguns fabricantes. </PT-narr>
</top>

<top>
<num> C221 </num>

<PT-title> Jogos Olímpicos de Inverno de 2002 </PT-title>
<PT-desc> Encontrar documentos sobre a selecção da cidade-sede dos Jogos Olímpicos de Inverno de 2002. </PT-desc>
<PT-narr> Documentos relevantes devem mencionar o processo de selecção e os nomes das cidades candidatas; documentos que mencionem apenas a cidade escolhida não são relevantes. </PT-narr>
</top>

<top>
<num> C222 </num>
<PT-title> Eleições presidenciais em França </PT-title>
<PT-desc> Encontrar documentos sobre quem ganhou e quem perdeu a segunda volta das eleições presidenciais francesas de 1995. </PT-desc>
<PT-narr> Documentos relevantes devem citar o nome da pessoa que ganhou a segunda volta das eleições presidenciais em Maio de 1995, mas também o nome do candidato derrotado nas mesmas eleições. </PT-narr>
</top>

<top>
<num> C223 </num>
<PT-title> Consequências de Chernobyl fora da antiga União Soviética </PT-title>
<PT-desc> Encontrar documentos sobre os efeitos da catástrofe nuclear de Chernobyl fora das antigas fronteiras soviéticas e as recomendações feitas numa conferência internacional dez anos após o desastre. </PT-desc>
<PT-narr> Documentos relevantes devem referir a conferência que teve lugar em Maio de 1995 para medir as consequências do acidente de Chernobyl. Documentos que mencionem os efeitos do desastre de Chernobyl fora da antiga União Soviética também são relevantes. </PT-narr>
</top>

<top>
<num> C224 </num>
<PT-title> Subida a solo feminina do Evereste </PT-title>
<PT-desc> Quem foi a primeira mulher a subir o Evereste sozinha e sem oxigénio? </PT-desc>
<PT-narr> Documentos relevantes devem citar o nome e/ou nacionalidade da primeira alpinista que conquistou o cume do Evereste sozinha e sem auxílio de oxigénio artificial. </PT-narr>
</top>

<top>
<num> C225 </num>
<PT-title> Central Nuclear de Sosnovyi Bor </PT-title>
<PT-desc> Encontrar documentos relatando o encerramento de emergência da central nuclear de Sosnovyi Bor em 1995. </PT-desc>
<PT-narr> Documentos relevantes devem descrever os problemas da central nuclear de Sosnovyi Bor, perto de São Petersburgo, em 1995. São de particular interesse documentos mencionando problemas com a segunda unidade. </PT-narr>
</top>

<top>
<num> C226 </num>
<PT-title> Operações de mudança de sexo </PT-title>
<PT-desc> Encontrar documentos relatando operações de mudança de sexo para transsexuais. </PT-desc>
<PT-narr> Todos os documentos que mencionarem operações de mudança de sexo para transsexuais são relevantes. </PT-narr>
</top>

<top>
<num> C227 </num>
<PT-title> A Donzela de Gelo do Altai </PT-title>
<PT-desc> Encontrar documentos relatando sobre a chamada "Donzela do Gelo do Altai" ou "Princesa do Gelo", uma mulher mumificada com mais de dois mil anos. </PT-desc>
<PT-narr> O corpo mumificado de um mulher descrita com uma donzela do gelo foi encontrado nas montanhas do Altai. O corpo foi conservado no solo "permafrost" da Sibéria por mais de 2000 anos. </PT-narr>
</top>

<top>
<num> C228 </num>
<PT-title> Arte pré-histórica </PT-title>
<PT-desc> Encontrar documentos sobre recentes descobertas de arte rupestre </PT-desc>
<PT-narr> Documentos relevantes devem citar informação sobre a localização de achados recentes de arte pré-histórica. </PT-narr>
</top>

<top>
<num> C229 </num>
<PT-title> Construção de Barragens </PT-title>
<PT-desc> Encontrar documentos sobre a construção de barragens. </PT-desc>
<PT-narr> Documentos relevantes devem fornecer alguns detalhes sobre a construção ou planeamento de barragens artificiais. Documentos meramente mencionando o nome ou localização de uma barragem não são relevantes. </PT-narr>
</top>

<top>
<num> C230 </num>
<PT-title> Atracagem Atlantis-Mir </PT-title>
<PT-desc> Encontrar documentos relatando a primeira atracagem entre o vaivém americano Atlantis e a estação espacial Mir. </PT-desc>
<PT-narr> Documentos relevantes devem mencionar a atracagem entre a Atlantis e a Mir. O número, nacionalidade e os nomes dos astronautas também são de interesse. </PT-narr>
</top>

<top>
<num> C231 </num>
<PT-title> Novo primeiro-ministro português </PT-title>
<PT-desc> Encontrar informação sobre as eleições legislativas em Portugal em Outubro de 1995 e o nome do recém-eleito primeiro-ministro. </PT-desc>
<PT-narr> Documentos relevantes devem mencionar especificamente as eleições em Portugal em 1995 e devem também citar o nome do novo chefe do Governo. </PT-narr>
</top>

<top>
<num> C232 </num>
<PT-title> Esquemas de pensões na Europa </PT-title>
<PT-desc> Encontrar documentos que descrevem os sistemas nacionais de pensões actualmente adoptados na Europa quer para nações individuais quer para todos os países da União Europeia. </PT-desc>
<PT-narr> Documentos relevantes devem conter informação sobre planos de pensão para estados europeus individuais ou para a União Europeia no seu conjunto. Informação de interesse incluiu idades mínimas e máximas para a reforma, o modo como o valor da pensão de reforma é calculado, tipos de pensão, percentagem de contribuições, etc. Planos para reformas futuras não são relevantes. </PT-narr>
</top>

<top>
<num> C233 </num>
<PT-title> Efeito de estufa </PT-title>
<PT-desc> Encontrar documentos sobre mudanças climáticas a nível global, e em particular discussões sobre o "efeito de estufa". </PT-desc>
<PT-narr> Documentos relevantes devem mencionar o efeito de estufa em relação às mudanças climáticas e o fenómeno do aquecimento global. Tanto documentos que argumentem que o efeito de estufa e o aquecimento global são a mesma coisa quer como documentos que o neguem são relevantes. </PT-narr>
</top>

<top>
<num> C234 </num>
<PT-title> Os surdos e a sociedade </PT-title>

<PT-desc> Encontrar informação sobre problemas encontrados por surdos na sociedade. </PT-desc>
<PT-narr> Documentos relevantes devem relatar qualquer assunto que afecte o bem-estar dos deficientes auditivos na vida social. </PT-narr>
</top>

<top>
<num> C235 </num>
<PT-title> Caça às focas </PT-title>
<PT-desc> Encontrar documentos discutindo a caça às focas e em especial as opiniões do WWF e outras organizações para a conservação da natureza. </PT-desc>
<PT-narr> Documentos descrevendo opiniões individuais sobre a caça às focas também são relevantes. </PT-narr>
</top>

<top>
<num> C236 </num>
<PT-title> Tufão nas Filipinas </PT-title>
<PT-desc> Qual o nome do furacão que se abateu sobre as Filipinas em Novembro de 1995, causando centenas de mortos e destruindo ou danificando milhares de edifícios. </PT-desc>
<PT-narr> Documentos relevantes devem mencionar o nome do tufão. Não é necessária mais informação. </PT-narr>
</top>

<top>
<num> C237 </num>
<PT-title> Panchen Lama </PT-title>
<PT-desc> Encontrar relatórios sobre as disputas relativas à selecção do novo Panchen Lama. </PT-desc>
<PT-narr> Documentos relevantes devem tratar das nomeações rivais relativas à reincarnação do Panchen Lama feitas pelo Dalai Lama e pelo governo da República Popular da China. Devem ser feitas referências ao facto que a aceitação do líder espiritual é contestada. </PT-narr>
</top>

<top>
<num> C238 </num>
<PT-title> Lady Diana </PT-title>
<PT-desc> Documentos relevantes devem detalhar as declarações feitas pela Princesa Diana sobre o seu casamento durante a sua famosa entrevista na BBC com Martin Bashir. </PT-desc>
<PT-narr> Documentos relevantes devem relatar afirmações feitas por Diana (e não por outros) relativas às dificuldades do seu casamento e às razões do fracasso deste </PT-narr>
</top>

<top>
<num> C239 </num>
<PT-title> Saúde mental dos jovens </PT-title>
<PT-desc> Encontrar documentos descrevendo problemas de saúde mental de crianças e adolescentes e possíveis tratamentos para esses problemas. </PT-desc>
<PT-narr> Todos os documentos que discutam de algum modo problemas de saúde mental de crianças ou adolescentes são relevantes. </PT-narr>
</top>

<top>
<num> C240 </num>
<PT-title> Camisa Fantasma dos Sioux </PT-title>
<PT-desc> Deve a camisa fantasma sagrada, em exposição num museu de Glasgow, ser devolvida à tribo norte-americana dos Sioux? </PT-desc>
<PT-narr> Negociações a decorrer entre responsáveis pelo museu de Glasgow e um grupo de índios tratam da devolução de uma camisa fantasma sagrada do povo Sioux roubada de uma vítima do massacre de Wounded Knee. Qualquer informação sobre essas negociações é relevante. </PT-narr>
</top>

<top>
<num> C241 </num>
<PT-title> Novos partidos políticos </PT-title>
<PT-desc> Encontrar documentos fornecendo informação sobre a fundação de novos partidos políticos, dando os nomes dos seus líderes. </PT-desc>
<PT-narr> Documentos relevantes devem discutir a fundação de novos partidos políticos e os nomes de pelo menos alguns dos seus líderes. Também é de interesse o país ou região onde o partido actua. Documentos sobre a refundação de partidos anteriormente existentes também são relevantes. </PT-narr>
</top>

<top>
<num> C242 </num>
<PT-title> Permanência recorde no espaço </PT-title>
<PT-desc> Encontrar documentos sobre a mais longa estadia de um ser humano no espaço, incluindo o nome do cosmonauta. </PT-desc>
<PT-narr> Documentos relevantes devem tratar do novo recorde para a mais longa estadia no espaço, mencionando o nome do cosmonauta que o obteve. </PT-narr>
</top>

<top>
<num> C243 </num>
<PT-title> Filmes de Kieslowski </PT-title>
<PT-desc> Encontrar documentos sobre os filmes do realizador polaco Krzysztof Kieslowski. </PT-desc>
<PT-narr> Documentos devem mencionar informação sobre os filmes de Kieslowski; a mera menção dos seus nomes não é suficiente. </PT-narr>
</top>

<top>
<num> C244 </num>
<PT-title> Futebolista do Ano 1994 </PT-title>
<PT-desc> Encontrar documentos que relatam a selecção do jogador mundial do ano da FIFA em 1994. </PT-desc>
<PT-narr> Em 1994 o futebolista do ano foi escolhido por jornalistas de 100 países em Lisboa. Documentos relevantes devem mencionar o nome desse jogador e fornecer informação sobre o processo de selecção ou os subsequentes festejos em Lisboa ou no seu país natal. Documentos que apenas mencionem o nome do vencedor sem mais detalhes não são relevantes. A mera menção do galardão noutros contextos não é suficiente. </PT-narr>
</top>

<top>
<num> C245 </num>
<PT-title> Christopher Reeve </PT-title>
<PT-desc> Encontrar documentos sobre a carreira do actor Christopher Reeve e o acidente que o paralisou. </PT-desc>
<PT-narr> Documentos relevantes devem tratar do actor, da sua carreira e do acidente que o paralisou. </PT-narr>
</top>

<top>
<num> C246 </num>
<PT-title> Castro visita a ONU </PT-title>
<PT-desc> Encontrar documentos relatando o discurso histórico de Fidel Castro na Assembleia Geral das Nações Unidas em 1995. </PT-desc>
<PT-narr> Documentos relevantes darão alguma informação sobre o conteúdo do discurso feito por Fidel Castro na Assembleia Geral das Nações Unidas em 1995. </PT-narr>
</top>

<top>
<num> C247 </num>
<PT-title> Túmulo de Alexandre Magno </PT-title>

<PT-desc> Encontrar detalhes sobre a descoberta da presumível localização do túmulo de Alexandre Magno. </PT-desc>
<PT-narr> Documentos relevantes devem relatar informação sobre a localização aproximada do local onde arqueólogos encontraram o que parece ser o túmulo de Alexandre o Grande. </PT-narr>
</top>

<top>
<num> C248 </num>
<PT-title> Disputa sobre o nome Macedónia </PT-title>
<PT-desc> Encontrar documentos sobre as objecções do governo grego relativas ao uso do nome Macedónia por uma das repúblicas da antiga Jugoslávia. </PT-desc>
<PT-narr> Documentos relevantes devem mencionar o desacordo e fornecer informação sobre a posição do governo grego e/ou as opiniões da população grega. </PT-narr>
</top>

<top>
<num> C249 </num>
<PT-title> Campeã dos 10.000 metros femininos </PT-title>
<PT-desc> Quem venceu os 10.000 metros femininos nos Mundiais de Atletismo em Gotemburgo? </PT-desc>
<PT-narr> Documentos relevantes devem nomear a vencedora da final dos dez mil metros nos Mundiais de Atletismo em Gotemburgo. </PT-narr>
</top>

<top>
<num> C250 </num>
<PT-title> Raiva em seres humanos </PT-title>
<PT-desc> Encontrar documentos relatando incidentes de raiva em seres humanos e discutindo métodos de prevenção da raiva em humanos. </PT-desc>
<PT-narr> Documentos relevantes devem permitir ao leitor informar-se sobre um ou mais métodos usados para a prevenção da forma humana da raiva. No entanto, documentos relatando incidentes de raiva em seres humanos também são de interesse. </PT-narr>
</top>

Anexo 2 – Os primeiros documentos recuperados pelo RDoc para o tópico 201

```
<DOCID>PUBLICO-19950407-087</DOCID>
<DATE>19950407</DATE>
<CATEGORY>Local</CATEGORY>
<AUTHOR>ACP</AUTHOR>
<TEXT>
Seixal
Padre suspeita de fogo posto
O pároco de Corroios admitiu ontem ter suspeitas de que o incêndio que
destruiu na terça-feira a Igreja de Vale de Milhaços tenha sido causado por
fogo posto. O padre Alberto Lino escusou-se a precisar as suspeitas, mas
relaciona o incêndio da igreja do Seixal com o que no domingo se registou
na capela de um lar de terceira idade em Vale de Figueira, Almada.
«Tenho as minhas suspeitas, mas como as bases não são sólidas não me atrevo
a expressá-las», afirmou o pároco, citado pela Lusa, esclarecendo que o
incêndio em Vale de Milhaços terá causado um prejuízo de cerca de 20 mil
contos e está a ser investigado pela PJ, que já procedeu a inquéritos.
Como justificação para a sua tese de fogo posto, Alberto Lino referiu o
facto de, no rescaldo do incêndio, não terem sido encontrados vestígios de
uma imagem em madeira de Nossa Senhora de Fátima, com 1,20 metros, comprada
há quatro meses por 280 contos.
</TEXT>
</DOC>
```

<DOC>
<DOCNO>PUBLICO-19950114-088</DOCNO>
<DOCID>PUBLICO-19950114-088</DOCID>
<DATE>19950114</DATE>
<CATEGORY>Local</CATEGORY>
<AUTHOR>ABRR</AUTHOR>
<TEXT>
Morte no trabalho
Um homem de 30 anos morreu, anteontem, vítima de acidente de trabalho, no lugar de Pera, concelho de Castanheira de Pera.
De acordo acordo com fontes da GNR local, Cid Manuel Gomes encontrava-se, pouco depois das 17 horas, a cortar madeira com uma moto-serra, instrumento que, por eventual descuido, o atingiu mortalmente. O cadáver foi removido para a morgue daquela vila do norte do distrito de Leiria, para ser autopsiado.
Entretanto, uma hora depois, no lugar de Eira da Pedra, freguesia de Maçãs de Caminho, concelho de Alvaiázere, morreu queimado, na lareira da sua residência, Angelo Amado Carvalho, de 35 anos, provavelmente por se ter deixado adormecer, caindo sobre o fogo. J.M.C.
</TEXT>
</DOC>
<DOC>

<DOCNO>PUBLICO-19950909-174</DOCNO>
<DOCID>PUBLICO-19950909-174</DOCID>
<DATE>19950909</DATE>
<CATEGORY>Nacional</CATEGORY>
<AUTHOR>CC</AUTHOR>
<TEXT>
Antecâmara silenciosa
«Eles estão mais para lá que para cá...» A frase ouviu-se à entrada do Lar de Idosos de Aljustrel quando os três principais candidatos da CDU pelo círculo de Beja distribuíam o seu manifesto eleitoral. Logo à entrada, dois locatários sentados receberam o manifesto, sem proferir uma palavra. Lá dentro o ambiente não foi mais animador. Rostos encarquilhados, com as marcas de uma vida inteira de privações, sem réstia de esperança no olhar. Quando o grupo entrou, ouviu-se a voz de uma mulher, numa cadeira de rodas, que perguntava «se vinha alguém que desse dinheiro». Não. A tarefa dos visitantes era outra. De qualquer modo, o Lar de Idosos de Aljustrel, instalado na antiga casa do administrador da companhia mineira, é um bom trunfo para José Soeiro, o cabeça de lista da CDU por Beja, pedir o voto às famílias dos velhos ali alojados. Afinal, foi a iniciativa do PCP em Aljustrel que criou o Lar, gerido agora pela Misericórdia mais pobre de todo o Alentejo. «Estão aqui idosos em estado terminal, à espera que a morte chegue. Pelo menos são acompanhados e não se tentam por um baraço e um ramo de uma qualquer velha azinheira.»
Manter o Lar é uma luta, diz o director da Misericórdia. Na antiga casa de família dos administradores da companhia mineira, conseguem-se instalar, com habilidade, 40 camas. A casa está hipotecada à Segurança Social, por dívidas. Por isso não se lhe pode mexer. A Câmara (CDU) quer ajudar, mas não é permitido fazer obras de ampliação. O candidato José Soeiro tomou apontamentos quando o seu número dois, António João Machado, funcionário técnico da Segurança Social, atirou com uma solução possível. A Segurança Social [Governo] negoceia com a empresa mineira, proprietária, e recebe as instalações como cobrança de parte da dívida. É uma solução para considerar logo depois das eleições, seja qual for o partido ganhador.
A encarregada do Lar contou um episódio ocorrido naquele mesmo dia. Um dos idosos residentes morreu às cinco e meia da manhã. Às oito, ainda a cama não tinha arrefecido, já o hospital local insistia para que o Lar recebesse outro idoso, em estado terminal. Este ambiente deprimente suscitou

comentários entre os candidatos a deputados. «Velhos abandonados?», perguntava-se. A resposta é simples. Os novos emigraram, para as cidades e para o estrangeiro. C.C.
</TEXT>
</DOC>