Indução de Filtros Linguisticamente Motivados na Recuperação de Informação

João Marcelo Azevedo Arcoverde

Indução de Filtros Linguisticamente Motivados na Recuperação de Informação

João Marcelo Azevedo Arcoverde

Orientadora: ***Profa. Dra. Maria das Graças Volpe Nunes***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional.

"VERSÃO REVISADA APÓS A DEFESA"

Data da Defesa: **17 / 04 / 2007**

Visto do orientador: _____

USP – São Carlos

Junho/2007

Agradecimentos

Aos meus pais, Carlos Fernando e Maria Alice, por conseguirem vencer imensos desafios, mostrando-me um caminho encorajador, de amor e doação. Acreditem, vocês fizeram o melhor por mim. Dedico-lhes esta obra.

À minha esposa, Simone, pelo companheirismo, compreensão e carinho. Os sentimentos que cultivamos a cada dia inspiram a evolução de nossas vidas. Estamos sempre aprendendo juntos.

À minha orientadora, Graça Nunes, por iluminar os caminhos e pela confiança em mim depositada. Amadureci bastante trabalhando ao seu lado. Agradeço sua dedicação, paciência e amizade.

Às Professoras Vera Lúcia Strube de Lima, da PUC-RS, e Renata Vieira, da Unisinos, por terem incentivado o intercâmbio entre nossas universidades, enriquecendo esta pesquisa.

Ao Professor Ricardo Baeza-Yates, por compartilhar intensas sinapses durante o mestrado. Aprendi que as soluções mais simples são as mais apropriadas.

Aos colegas da USP, pelo precioso tempo em que estivemos juntos. Meus sinceros agradecimentos pela disponibilização de conhecimento relevante.

Ao meu sócio, André, pela honestidade, amizade e força. A prosperidade não vem ao acaso. É fruto do incessante trabalho que temos desenvolvido. Enfrentamos juntos esta jornada.

A todos os meus amigos, de todas as épocas, da imensa São Paulo e da minha terra natal, Recife, que influenciaram minha personalidade e me ajudaram a buscar a felicidade.

A Deus e todas as suas manifestações, em especial pelo milagre da Vida.

Resumo

Apesar dos processos de recuperação e filtragem de informação sempre terem usado técnicas básicas de Processamento de Linguagem Natural (PLN) no suporte à estruturação de documentos, ainda são poucas as indicações sobre os avanços relacionados à utilização de técnicas mais sofisticadas de PLN que justifiquem o custo de sua utilização nestes processos, em comparação com as abordagens tradicionais. Este trabalho investiga algumas evidências que fundamentam a hipótese de que a aplicação de métodos que utilizam conhecimento linguístico é viável, demarcando importantes contribuições para o aumento de sua eficiência em adição aos métodos estatísticos tradicionais. É proposto um modelo de representação de texto fundamentado em sintagmas nominais, cuja representatividade de seus descritores é calculada utilizando-se o conceito de "evidência", apoiado em métodos estatísticos. Filtros induzidos a partir desse modelo são utilizados para classificar os documentos recuperados analisando-se a relevância implícita no perfil do usuário. O aumento da precisão (e, portanto, da eficácia) em sistemas de Recuperação de Informação, conseqüência da pós-filtragem seletiva de informações, demonstra uma clara evidência de como o uso de técnicas de PLN pode auxiliar a categorização de textos, abrindo reais possibilidades para o aprimoramento do modelo apresentado.

Palavras-chaves: Recuperação de Informação, Filtragem de Informação, Processamento de Linguagem Natural, Sintagmas Nominais, Categorização de Textos, Aprendizado de Máquina

Abstract

Although Information Retrieval and Filtering tasks have always used basic Natural Language Processing (NLP) techniques for supporting document structuring, there is still space for more sophisticated NLP techniques which justify their cost when compared to the traditional approaches. This research aims to investigate some evidences that justify the hypothesis on which the use of linguistic-based methods is feasible and can bring on relevant contributions to this area. In this work noun phrases of a text are used as descriptors whose evidence is calculated by statistical methods. Filters are then induced to classify the retrieved documents by measuring their implicit relevance presupposed by an user profile. The increase of precision (efficacy) in IR systems as a consequence of the use of NLP techniques for text classification in the filtering task is an evidence of how this approach can be further explored.

Keywords: Information Retrieval, Information Filtering, Natural Language Processing, Noun Phrases, Text Categorization, Machine Learning

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

AM Aprendizado de Máquina

CT Categorização de Textos

FI Filtragem de Informação

FLM Filtro Linguisticamente Motivado

IA Inteligência Artificial

IC Indução Construtiva

MT Mineração de Textos

PLN Processamento de Linguagem Natural

RI Recuperação de Informação

SN Sintagma Nominal

1 Introdução

1.1 Contexto

A Web tem se revelado um fenômeno irreversível, em vista de seu exponencial crescimento e de sua capacidade de compartilhar conhecimento, incrementar comunicação e desenvolver novas formas de comércio. Sua riqueza e diversidade de informações tornam-na o maior e o mais importante repositório de informação da História. Entretanto, sua natureza, estrutura e dimensão dificultam sua manipulação de forma sistemática e eficiente pelos diferentes tipos de usuários.

A velocidade de mudança a que hoje os fenômenos sociais, inovações científicas, tecnológicas e culturais estão suscetíveis, e que nos desperta ansiedade pela busca incessante de informação atualizada, faz da Web o mais importante vetor de disseminação de novos conhecimentos, frente a todas as outras mídias, como rádio, TV, livros e jornais. Seu poder de ofertar acesso às fontes de informações em tempo quase real é caracterizado por uma maneira desorganizada, incontrolável e de difícil manipulação. Sua infometria corresponde ao produto do trabalho colaborativo de milhões de pessoas cujos esforços, na maioria das vezes, são mínimos, representando de forma exemplar o Princípio do Esforço Mínimo (Zipf, 1949).

No seu vasto universo de possibilidades, estamos praticamente limitados ao volume de informações que conseguimos digerir, ou ao menos perceber. Somos então constantemente desafiados à árdua tarefa de filtrar aquilo que vai ao encontro de nossos interesses imediatos ou de caráter mais duradouro, sendo difícil encontrar informação relevante, principalmente porque há muita informação irrelevante na Web (Baeza-Yates e Ribeiro-Neto, 1999), ainda que relevância seja um conceito situacional, subjetivo e dependente de contexto.

Saber o que se quer buscar e, não menos importante, saber como buscar, tornam-se requisitos imprescindíveis para o sucesso da recuperação eficiente de informação. As ferramentas hoje disponíveis para essas atividades são os principais artefatos funcionais para nos auxiliar quando sabemos o que queremos. Usá-las eficientemente é um desafio até para os mais experientes internautas. É certo que a Web possui muitas respostas para perguntas nunca imaginadas (Baeza-Yates, 2004b) e, por muitas vezes, em situações específicas, o

modelo de busca por informação relevante através destas ferramentas convencionais como *Google*[1], *Yahoo*[2] e *AllTheWeb*[3], por exemplo, não atende adequadamente às necessidades do usuário, seja pela inabilidade desse em saber expressar suas intenções, ou seja pelo grau de desenvolvimento tecnológico em que essas ferramentas se encontram ao trabalhar cada etapa do ciclo iterativo de um sistema complexo como a Web.

A dinâmica destes desafios é o campo de estudos da Gerência do Conhecimento, que utiliza como apoio instrumentos computacionais que permitem às organizações capturar, organizar, compartilhar e otimizar a utilização do conhecimento (Wilcox, 1997). O enfoque maior destas áreas tem sido em métodos, modelos e técnicas para auxiliar a sistematização do tratamento da informação disponível ao alcance do usuário. Contudo, pesquisas sobre como lidar com a sobrecarga de informação na Web de forma a extrair o máximo de benefícios de seu conteúdo ainda está em seu início, como por exemplo, mecanismos de recuperação que garantam a qualidade da informação retornada sob altos padrões de precisão e revocação.

Pesquisas sobre como lidar com a sobrecarga de informação na Web de forma a extrair o máximo de benefícios de seu conteúdo ainda são necessárias, como por exemplo, para os atuais mecanismos de recuperação garantirem a qualidade da informação retornada com altas precisão e revocação. A partir dos anos 80 registrou-se o aumento intensivo dos processos exploratórios de descoberta, representação e processamento do conhecimento através da Inteligência Artificial (IA), juntamente com o auxílio de outras áreas de pesquisa como a Estatística e a Linguística, através de Sistemas Especialistas ou Sistemas Baseados em Conhecimento (Tsui et al., 2000), que muito contribuíram neste sentido. Todavia, o avanço de técnicas de Mineração de Textos (MT), Processamento de Linguagem Natural (PLN), Aprendizado de Máquina (AM), entre outras sub-áreas de pesquisas interdisciplinares da IA, está possibilitando resultados mais otimistas para certas classes de problemas relacionados à sobrecarga da informação.

[1] www.google.com
[2] www.yahoo.com
[3] www.alltheweb.com

1.2 Motivação e relevância

Tem-se presenciado uma procura contínua por abordagens alternativas para a representação estruturada de textos que, muitas vezes, focam a utilização de modelos híbridos, como é o caso da Recuperação de Informação (RI) e Filtragem de Informação (FI) com motivação linguística e apoiados em métodos estatísticos. Nesses modelos híbridos, a representação do conteúdo dos documentos começa a usufruir de técnicas mais avançadas de PLN que permitem explorar as características coesivas intrínsecas aos textos analisados (Sparck-Jones e Willett, 1997; Dias et al., 2000), com relativo sucesso em comparação aos métodos tradicionais no tocante ao aumento de precisão e revocação para sistemas de RI (Gonzalez, 2005).

Apesar dos processos de RI sempre terem usado técnicas básicas de PLN no suporte à estruturação de documentos em tabelas atributo-valor (atomização, remoção de stopwords, lematização, radicalização, conflação, nominalização, etc), ainda são poucas as indicações sobre os avanços relacionados à utilização de técnicas mais sofisticadas de PLN que justifiquem o custo de sua utilização nesses sistemas, frente aos benefícios observados (Smeaton, 1999). Esse trabalho propõe a investigação de algumas evidências que fundamentam a hipótese de que certas tarefas de PLN aplicadas em domínios específicos demarcam importantes contribuições para o aumento da eficiência dos sistemas de RI, em adição aos métodos estatísticos tradicionais.

Existem dois grupos de estratégias de RI quanto ao modelo adotado para representação de consultas e documentos: as que usam modelos unigrama (amplamente conhecidos como *bag-of-words*) e as que usam modelos com dependência de termos (Gonzalez, 2005).

Uma suposição fundamental adotada pelas abordagens mais comuns de RI (vetorial e probabilística) é a que assume a independência entre os termos que constituem os documentos (Robertson e Sparck-Jones, 1976; Robertson, 1977; Salton, Buckley e Yu, 1982; Cooper, 1995). As associações entre os termos do texto, necessárias para atribuir-lhe valor semântico, não são reconhecidas, bem como qualquer mecanismo de coesão frásica. A suposição da independência entre os termos é conveniente por facilitar e minimizar o número de parâmetros que precisam ser estimados nessas abordagens tradicionais (Losee, 1989).

É reconhecido que o cálculo do peso da representatividade dos atributos do texto pode ser determinado mais facilmente se eles ocorrerem de forma independente uns dos

outros. Pode-se computar o peso de cada atributo de forma independente, viabilizando estatisticamente o processo. Todavia, realizar a indexação de documentos utilizando somente métodos estatísticos fundamentados em modelos unigrama compromete a eficácia em sistemas de RI, fato percebido desde o fim da década de 80 (Smeaton, 1990).

Alternativamente, é possível a utilização de padrões linguísticos resultantes da dependência dos termos na representação estruturada dos textos para aplicações que utilizem PLN em seu processo exploratório, buscando-se aumentar seus níveis de eficácia. Contudo, existe um preço a ser pago na utilização do PLN nessas aplicações, o que às vezes pode inviabilizar seu uso em sistemas reais de RI, por exemplo, onde o tempo de resposta é um dos critérios decisivos para avaliação.

1.3 Objetivos

Considerar um caso especial de dependência de termos na representação de documentos através de um modelo linguisticamente motivado, a fim de que se possa avaliar seu impacto nos resultados de um processo de RI - auxiliado por um subsistema de FI, em adição a modelos unigrama tradicionalmente utilizados, constitui o objetivo desta proposta.

O presente trabalho propõe-se a adotar o conceito de descritores como unidade básica de indexação. Descritores são termos portadores de informação e que fazem referência a um conceito, objeto ou fato do mundo real. Portanto, os elementos que devem ser extraídos de um documento para representá-lo devem possuir a mesma função de um descritor. Entre as variedades de relacionamentos mais encontradas em modelos com dependência de termos, adotamos o Sintagma Nominal (SN), que representa a menor parte do discurso portadora de informação (Kuramoto, 2002).

Pretende-se também investigar formas de calcular o peso desses descritores no modelo com dependência de termos fundamentado em sintagmas nominais. Isto é, quais são as alternativas para a computação dos pesos que os descritores representam no texto ? Modelos tradicionais estão fundamentados na frequência de ocorrência das palavras nos documentos (Salton e Buckley, 1987a). Entretanto, alguns trabalhos mais recentes incorporam, também, informação morfológica (Vilares et al., 2002) ou sintática (Lee e Lee, 2005) para compor o espaço de descritores. A combinação de abordagens estatísticas e

linguísticas parece indicar o caminho mais promissor (Sebastiani, 2002), e esta tendência será considerada e avaliada neste trabalho.

Foi utilizado o conceito de "evidência" proposto por Pickens e Croft (2000), de maneira a contribuir para o cálculo da representatividade destes descritores. Especulam-se, no modelo composto por sintagmas nominais como unidades básicas de indexação, alternativas para a combinação do conceito de evidência com as formas tradicionais de calcular a representatividade dos descritores, buscando-se melhorar a classificação de relevância dos documentos.

Para avaliar os resultados empíricos sob a ótica do impacto do modelo de representação proposto na precisão dos sistemas de RI, convenciona-se utilizar a saída destes sistemas para alimentar um subsistema de FI. A Web, por ser um imenso repositório de informações, possibilita que o resultado da busca produzida por uma simples consulta possa servir como entrada para um sistema de filtragem de informações. Uma das principais características da Web é o dinamismo evolutivo da sua coleção de documentos, que é praticamente desconhecida ou muito pouco conhecida, diferentemente do que ocorre em sistemas estáticos de RI, que possibilitam aos engenhos de busca atingir altos níveis de revocação e precisão.

De forma análoga, as consultas em um sistema tradicional de RI são praticamente imprevisíveis e voláteis, pois representam situações instanciais dos usuários do sistema. Já em um sistema de FI, elas sugerem perfis, que são representações mais "duradouras" dos interesses desses usuários, que filtram seletivamente o que satisfaz suas necessidades dentre todo o volume de informação que tenha sido gerado. Conceitualmente, um sistema de FI diz respeito ao processo de acesso e recuperação de informação em bancos de dados remotos, no qual os dados para análise são produto direto da busca nesta base através de um sistema de RI (Belkin e Croft, 1992).

O processo de FI pode ser descrito como um problema especial de classificação binária, e sua principal atividade é a classificação de documentos (advindos de um fluxo) em relevantes ou irrelevantes, em função do interesse particular do usuário (mantido pelo seu "perfil") com o objetivo de reduzir a sobrecarga de informação. Consequentemente, essa proposta de trabalho também pretende avaliar como o modelo linguisticamente motivado baseado em sintagmas nominais influencia os algoritmos proposicionais de AM, tendo em vista que a tabela atributo x valor que representa a estrutura dos documentos no processo

de aprendizado é a mesma que representa o espaço de descritores no processo de indexação do sistema de RI.

1.4 Organização da dissertação

Essa dissertação encontra-se dividida em oito capítulos, a seguir:

No Capítulo 2 é consolidada a revisão bibliográfica de alguns dos trabalhos mais representativos e recentes da área de FI e representação de documentos que levam em consideração a dependência de termos.

No Capítulo 3 é apresentada nossa proposta de modelo estruturado de representação de documentos com dependência de termos, utilizando os sintagmas nominais como as principais unidades de indexação. São explicitadas as estratégias automáticas de extração dos SNs através de técnicas estatísticas e simbólicas, bem como é proposto um método para calcular a representatividade dos descritores baseado no conceito de evidência.

No Capítulo 4 é descrito o processo de Filtragem de Informação (FI) como uma atividade de Categorização de Textos, e introduzido o conceito de Filtro Linguisticamente Motivado (FLM). São apresentandos os algoritmos de Aprendizado de Máquina mais comumente utilizados para o processo de FI, dentre os quais aqueles utilizados nesse experimento.

No Capítulo 5 é demonstrada uma experiência prática de expansão automática de consultas com análise local de sintagmas nominais, realizada para o *CLEF 2006* (*Cross-Language Evaluation Forum*[4]).

No Capítulo 6 é descrito um experimento de FI sobre um fluxo dinâmico de documentos retornados pelo sistema de RI. Sua arquitetura e funcionamento são detalhados, incluindo o papel fundamental dos julgamentos de relevantes utilizados para o processo de formação da base de treinamento dos filtros. O modelo híbrido de representação de textos que utiliza um espaço de descritores com dependência de termos é apresentado, bem como o processo de indução construtiva usado para determinar os descritores multitermos

[4] http://www.clef-campaign.org/

a partir dos sintagmas nominais identificados nos textos. Por fim, é feita uma avaliação do impacto desse modelo sobre alguns tópicos de consulta especialmente selecionados para o experimento.

O Capítulo 7 encerra as conclusões obtidas sobre toda a experiência, destacando algumas observações importantes sobre os processos, indicando as implicações e sugestões de trabalhos futuros.

2 Revisão bibliográfica

Devido à importância da identificação de trabalhos correlatos aos objetivos apresentados, foram selecionados alguns mais representativos para a modelagem da dependência de termos e para o processo de FI, de acordo com critérios que levam em conta a sua relação com a presente proposta e a atualidade da publicação.

O marco inicial da pesquisa pela viabilidade de métodos linguísticos na RI encontra-se registrado em Salton e McGill (1986), destacando-se a análise sintática dos textos para a identificação das estruturas sintagmáticas. Souza e Alvarenga (2004) apontam as dificuldades intrínsecas ao processo de análise semântica através da análise sintática e exemplificam casos em que é impossível o reconhecimento não ambíguo de relações semânticas através dos componentes da sentença, sugerindo que um modelo baseado em gramáticas poderia trazer melhores resultados.

Strzalkowski (1995) adotou um analisador sintático (*parser*) simbólico, baseado em uma gramática com regras mantidas manualmente e que proviam certas heurísticas para desambiguar a estrutura de um sintagma nominal no texto com relativo sucesso. Contudo, sua abordagem não provia escalabilidade e robustez para coleções de documentos acima de alguns *megabytes*. Em análises posteriores, Strzalkowski percebeu que o esquema *tf.idf*[5] (Salton e Buckley, 1987b) não era apropriado para representar o peso dos descritores de diferentes tipos combinados no mesmo modelo, a exemplo de termos simples + relacionamentos. Desta forma, os pesos dos relacionamentos foram acrescidos de parâmetros que representam fatores de multiplicação.

Zhai (1997) propõe um modelo de *parser* probabilístico, robusto o suficiente para indexar sintagmas nominais (SNs) em grandes coleções de documentos (em torno de 250Mb), provendo a possibilidade de combinar diferentes tipos de sub-frases (descritores) derivadas dos SNs identificados: termos, pares modificado-modificador e o próprio sintagma.

Chandrasekar e Srinivas (1997) propuseram o uso de técnicas linguisticamente motivadas no processo de RI, aumentando sua eficiência através da exploração de informações latentes nos textos. Utilizaram estruturas sintáticas e padrões de uso da linguagem extraídos por um etiquetador morfo-sintático para eliminar resultados irrelevantes, por meio

[5] Term Frequency x Inverted Document Frequency

de um filtro aplicado na fase de apresentação dos resultados em um sistema de RI. Seu trabalho de filtragem sintática atingiu precisão e revocação de 97% e 86%, respectivamente.

Pickens e Croft (2000), investigaram os sintagmas e suas propriedades independentemente de qualquer abordagem específica de RI, não apenas para um melhor entendimento destas estruturas mas também para explorar melhores métodos para analisá-los, determinando o valor de várias formulações frasais para o processo de RI. Os sintagmas são identificados e extraídos do texto usando uma abordagem estatística *markoviana*, produzindo melhores resultados na extração de SNs de alta qualidade, em comparação com métodos sintáticos de extração baseados em gramática. Neste trabalho, os autores defendem o uso do conceito de *evidência* como um artifício eficiente para determinar o peso dos descritores em modelos híbridos (linguísticos e estatísticos) de representação de documentos.

Mais recentemente, Liu et al. (2004) e co-autores trabalharam na identificação de SNs na consulta e recuperação de documentos utilizando a *WordNet* (Miller e Fellbaum, 1990, aput) para desambiguação dos termos da consulta. Essa estratégia considera que um documento é relevante para uma consulta com um determinado SN se todos os termos deste SN ocorrem dentro de uma janela de texto de tamanho específico para cada tipo de sintagma, utilizando uma árvore de decisão para o cálculo do tamanho dessa janela.

Cientistas brasileiros também revelam pesquisas sobre a utilização dos SNs em processos computacionais. Kuramoto (2002) defende a viabilidade de uso dos SNs como os principais descritores de documentos em processo de indexação, pelo maior grau de informação semântica embutida nestas estruturas. Em sua tese de doutorado, Kuramoto desenvolveu uma pesquisa fundamental para a consideração dos SNs como descritores em textos. Entretanto, na época em que foi desenvolvido seu trabalho, a extração dos SNs foi realizada de forma manual, simulando uma extração automática, pela ausência de ferramentas computacionais customizadas para o caso especial da língua portuguesa.

Miorelli (2001) apresenta um método de extração de SNs em sentenças da língua portuguesa para ser usado no processo de indexação em sistemas de RI. O trabalho apresenta um estudo detalhado da estrutura do SN para a identificação de regras que venham a compor uma gramática de SN. Esse método utiliza uma gramática criada manualmente, que descreve o comportamento sintático do SN para realizar a análise das sentenças. Conforme foi constatado em trabalhos similares (Voutilainen, 1993), a produção artesanal deste co-

nhecimento é muito custosa e, na maioria dos casos, não é compartilhável entre diferentes aplicações.

Gonzalez (2005) apresenta o processo de nominalização[6] com melhores resultados em comparação com a normalização lexical[7]. Explora ainda as Relações Lexicais Binárias (RLBs)[8] no processo de aquisição de informação linguística e o uso de consultas booleanas para contribuir na especificação de dependência de termos. Investiga o cálculo da representatividade dos descritores baseado em evidência, apresentando vantagens em relação ao cálculo baseado em frequência de ocorrência. Os experimentos relatados indicam que estes recursos melhoram os resultados de sistemas de RI.

Aires (2005) discorreu sobre a possibilidade técnica de classificar os resultados das buscas segundo os objetivos e intenções dos usuários. Ao invés da análise do conteúdo dos textos, foram escolhidas características linguísticas relacionadas com o estilo de documentos em português, e desenvolvidos estudos com usuários para avaliar os classificadores criados. A visão centrada no usuário em combinação com o uso de marcadores estilísticos para a indução dos classificadores demarca uma inovação nos métodos tradicionais utilizados para conseguir melhores resultados nos sistemas de RI para lidar com a sobrecarga de informação.

Santos (2005), em sua dissertação de mestrado, demonstrou o uso prático do algoritmo de AM denominado Aprendizado Baseado em Transformações (*TBL - Transformation Based Learning*) para a extração de sintagmas nominais para o caso do português brasileiro, guiado por um corpus de treino previamente classificado e revisado manualmente[9]. Foi verificado que, para a obtenção de um conjunto de SNs mais ricos em termos de conteúdo, foi necessária a identificação de SNs contendo os pós-modificadores adjetivos e sintagmas preposicionados. Tal abordagem é diferente da que tem sido frequentemente utilizada em trabalhos de língua inglesa, que normalmente usam o conceito de SNs básicos. O sistema atingiu precisão e abrangência ligeiramente superior a 90%.

[6] processo alternativo de normalização morfológica que deriva substantivos a partir de palavras de outras categorias gramaticais, principalmente verbos e adjetivos

[7] compreende o processo de conflação, *stemming* e lematização

[8] relacionamentos entre termos nominalizados que capturam mecanismos de coesão frásica

[9] Mac-Morpho, do NILC: http://www.nilc.icmc.usp.br/lacioweb/corpora.htm

O trabalho de Kjersti (1997) é uma clássica referência sobre o processo de FI. Ele destaca diversos sistemas acadêmicos e comerciais, abordando vários métodos de representação e indexação de textos para o processo de FI, além da Aquisição de Conhecimento sobre os interesses do usuário e a indução de *perfis* através de técnicas de Aprendizado de Máquina.

Em sua tese de doutorado, Baudisch (2001) trata a dinâmica da mudança de foco dos usuários ao longo de sua experiência de busca por informações. Nesse processo, o "perfil" através do qual é mantida a representação de seus interesses perde sua acurácia, causando a queda da qualidade preditiva do filtro na classificação de novos documentos. Quanto mais dinâmicos são os interesses dos usuários, mais importantes eles são para atualizar seus perfis, de maneira a encurtar o período de queda de sua qualidade preditiva. O objetivo de sua dissertação é conceber uma nova arquitetura de Filtragem de Informação capaz de suportar a frequente mudança dos interesses de seus usuários, através de uma estrutura modular denominada QSA (*QuerySet filtering architecture*). Essa arquitetura consiste em um conjunto de consultas que possui um alto nível de correspondência com cada um dos interesses do usuário. A mudança de interesse afeta apenas um subconjunto de seu perfil representativo, atualizando apenas aquelas consultas que correspondem aos interesses afetados.

O processo de FI lida com a monitoração de fluxos de textos para a deteção de padrões mais complexos do que as consultas (*queries*) tratadas pelos engenhos de busca. Neste sentido destacamos o trabalho *InfoFilter* (Elkhalifa et al., 2005), cujo foco está na especificação de padrões através de uma linguagem denominada *Psnoop (Pattern Specification Language)* e na sua respectiva deteção através do uso de um paradigma de fluxo de dados chamado *Pattern Detection Graph (PDG)*.

Estruturas de dados complexas como *árvores de sufixos* e técnicas de programação dinâmica também são usadas para a deteção de múltiplos padrões de textos de forma simultânea, por aproximação, em grandes coleções de documentos. Nesta linha salientamos o trabalho intitulado *Matchsimile* (Navarro, Baeza-Yates e Arcoverde, 2003), com aplicabilidade em diversas áreas da computação, desde bioinformática até processos de FI em bases textuais com alta suscetibilidade a erros tipográficos. Hoje o algoritmo proposto é largamente utilizado no Brasil na monitoração sistemática em Diários Oficiais para sistemas de filtragem e roteamento de informações.

Um estudo comparativo de métodos para calcular os pesos dos termos para o processo de FI foi apresentado por Nanas et al. (2003), destacando os diferentes requerimentos para aqueles métodos advindos da RI e da categorização de textos, respectivamente.

Para finalizar nossa análise bibliográfica, citamos duas recentes teses que lidam com filtros adaptativos, que sintonizam com a natureza dinâmica e mutável dos interesses do usuário, constituintes principais do seu perfil de busca. O primeiro deles chama-se *Nootropia* (Nanas, Uren e Roeck, 2004) e compreende um perfil multi-tópico do usuário através de uma rede hierárquica de termos que leva em consideração a topicalidade e correlações lexicais entre estes termos. A segunda tese (Macskassy, 2003) introduz um novo tipo de critério que compõe o interesse do usuário de maneira prospectiva, i.e., o critério define o grau de interesse de uma informação baseada em eventos que acontecem subsequentemente aos itens que aparecem em seu perfil de busca.

No próximo Capítulo trataremos do processo de estruturação de textos livres, apresentando os diversos modelos de representação de documentos mais comumente relacionados na literatura, com e sem dependência de termos considerada na estratégia do processo de estruturação.

3 Modelos de representação de documentos

O formato textual é a forma mais natural de armazenar informações e a manipulação computacional de textos constitui um processo fundamental para a automatização da captura, seleção, organização e compartilhamento de ativos do conhecimento. Devido à crescente disponibilidade de documentos digitais e à necessidade de acessá-los eficientemente, hoje a evolução desse processo constitui o principal desafio no campo de sistemas de informação, frente aos sistemas de gerenciamento de banco de dados relacionais. O tratamento computacional de textos é um processo complexo principalmente pelo formato não estruturado em que esses estão disponibilizados. Outros fatores também contribuem para o desafio do processamento de textos, tais como ambiguidade, idioma, estilo e domínio.

3.1 Estruturação de textos livres

Algumas atividades relacionadas ao processamento de textos livres - aqueles contidos em documentos desprovidos de estruturas que especifiquem seu conteúdo, não exigem necessariamente uma representação estruturada do texto para a sua manipulação direta. É o caso da busca por padrões de texto através de máquinas de estado finito, comparação de cadeias de caracteres para determinar a distância de edição[10] entre elas, criptografia, stenografia, compressão de textos, entre outras computações relacionadas à área de *stringologia* (Navarro e Raffinot, 2002). O processamento do texto é dito *online* quando for realizado de forma sequencial, analizando-se o texto caractere por caractere, convencionalmente da esquerda para a direita.

Entretanto, muitas outras atividades necessitam que a coleção de textos a ser trabalhada seja pré-processada com o intuito de representá-la através do mapeamento de seu conteúdo em um formato estruturado e compacto, de forma a atender diferentes necessidades de processamento. Esse processo chama-se indexação, e a estrutura de dados empregada na formação do índice pode variar conforme a aplicação, a necessidade de performance dos algoritmos empregados ou mesmo o espaço demandado pela sua representação. Sistemas de Aprendizado de Máquina, Filtragem e Recuperação de Informação usualmente necessi-

[10] número mínimo de mutações pontuais (inserção, remoção, substituição e transposição) requeridas para transformar uma cadeia em outra.

tam estruturar o texto livre para que se faça possível sua manipulação pelo conjunto de algoritmos empregados nestas aplicações.

Um problema fundamental ao lidar com a representação da linguagem natural é que o contexto tem uma influência substancial na interpretação do significado de uma passagem no texto. Diferentes abordagens de representação de texto podem considerar mais ou menos os fenômenos linguísticos, a exemplo da sinonímia e polissemia (descritos no Capítulo 5), tão problemáticos para modelos de representação de textos. As abordagens podem ser classificadas de acordo com o nível em que elas analisam o texto:

- Sub-palavra: decomposição das palavras e sua morfologia (morfemas e/ou n-grama);
- Palavra: palavras e informação léxica;
- Multi-palavras: frases e informação sintática
- Semântico: significado do texto
- Pragmático: significado do texto em relação ao contexto (ex: estrutura de diálogo)

Os blocos básicos em cada nível são chamados de termos de indexação. No nível das multi-palavras, por exemplo, os termos de indexação referem-se à expressões, frases ou sentenças inteiras. O caso dos sintagmas nominais, foco deste trabalho, é reconhecido como um caso especial de unidade de indexação onde existe uma dependência funcional entre os termos que compõem essas estruturas.

Apesar dos benefícios para a Linguística Computacional ao estruturar o processamento de linguagem natural nessas categorias, elas não podem ser tratadas de forma independente, pois em cada nível existem ambiguidades que só podem ser resolvidas no nível imediatamente seguinte. Por exemplo, para identificar se uma palavra é um substantivo ou um verbo quando ambos assumem a mesma forma, é necessário subir ao nível multi-palavras e verificar a informação sintática da frase em que a palavra se encontra.

De forma geral, quanto maior o nível, mais detalhes sobre o texto é possível capturar, mas também é maior a complexidade para produzir as representações automaticamente. O nível mais comum de representação de texto para tarefas de classificação é o da palavra, pois na maioria dos casos essas são unidades significativas de pouca ambiguidade, mesmo sem considerar o contexto, pois apesar de existirem palavras homógrafas, assume-se que elas

têm pouco impacto na representação do documento como um todo. A principal vantagem desse nível é a simplicidade de implementação do modelo de representação de textos.

Em geral, é desconsiderada a ordem com que as unidades de indexação aparecem no texto, simplificando o processo de representação de textos, considerando apenas a frequência com que essas unidades aparecem nos documentos, enquanto que toda a estrutura desses documentos é ignorada. Essa representação é conhecida por *bag-of-words*. Assim, linguagens de descrição são necessárias para representar o espaço compartilhado pelas unidades de indexação. Em geral, essas linguagens podem ser divididas em dois tipos: linguagem baseada em atributo-valor ou proposicional e linguagem relacional. Todo o nosso trabalho é baseado em modelos proposicioinais de indexação dos descritores, muito embora exista o emprego de conhecimento linguístico para a seleção e construção dos descritores.

3.2 Redução de dimensionalidade

O processo de representação de textos livres em um formato estruturado constitui a etapa mais importante de um sistema de AM ou de RI, pois influencia substancialmente o sucesso da aplicação dos algoritmos envolvidos na obtenção dos resultados almejados (Sebastiani, 2002). A escolha do modelo de representação de textos depende daquilo que é considerado como unidade significativa desses textos e das regras de linguagem natural adotadas para a combinação dessas unidades. Nem todas as palavras são igualmente significativas para representar a semântica de um documento, pois algumas palavras carregam mais significado que outras[11]. Segundo Smeaton (1997), é complexo o tratamento por sistemas de RI daqueles casos nos quais palavras diferentes são usadas para representar o mesmo significado ou conceito dentro dos documentos ou consultas. Analogamente, um sistema de RI não pode suportar facilmente palavras polissêmicas, que podem contemplar múltiplos significados.

Basicamente, um termo de índice ou unidade de indexação é uma palavra ou um conjunto de palavras relacionadas que possui um significado por si só, podendo ser qualquer palavra que apareça no documento. Modelos de representação de textos baseados puramente nesse conceito consideram que a semântica de um documento pode ser expressa

[11] Usualmente os substantivos constituem a classe gramatical mais representativa do conteúdo de um documento, seguidos de adjetivos e verbos, ou seja, as classes abertas.

simplesmente através de um conjunto de termos de índice. Claramente essa assertiva resume-se a uma simplificação do problema porque grande parte da semântica de um documento é perdida ao substituir seu texto por um simples conjunto de palavras (Baeza-Yates e Ribeiro-Neto, 1999). Dessa forma, em alguns sistemas de RI, os documentos recuperados em resposta a uma necessidade de consulta de um usuário são frequentemente irrelevantes.

Devido à forma não estruturada dos documentos, a etapa de indexação (ou pré-processamento) dos textos para um formato estruturado, tal como uma tabela atributo-valor, costuma ser a mais custosa em relação a outros processos envolvidos em um sistema de RI ou AM. Mesmo após todo o trabalho envolvido, a representação final de documentos em uma forma estruturada é caracterizada pela alta dimensionalidade e esparsividade do modelo adotado, sendo de fundamental importância a seleção criteriosa de quais termos serão considerados como atributos significativos para a sua representação. Essas características são inerentes a problemas relacionados ao processo de Mineração de Textos (MT)[12], pois cada palavra presente nos documentos pode ser um possível candidato ao conjunto de atributos dessa tabela atributo-valor, como ilustrado na Tabela 1.

	t_1	t_2	$\ldots$	$t_{\|T\|}$
d_1	w_{11}	w_{12}	$\ldots$	$w_{1\|T\|}$
d_2	w_{21}	w_{22}	$\ldots$	$w_{2\|T\|}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$d_{\|D\|}$	$w_{\|D\|1}$	$w_{\|D\|2}$	$\ldots$	$w_{\|D\|\|T\|}$

$$\textit{onde} \begin{cases} d_1 \text{ a } d_{|D|} \text{ são os documentos da coleção;} \\ t_1 \text{ a } t_{|T|} \text{ compõem o espaço de descritores;} \\ w_{11} \text{ a } w_{|D||T|} \text{ são os pesos relacionados a cada descritor.} \end{cases} \quad (1)$$

Tabela 1: Representação de documentos

Usar todo o conjunto de palavras disponível na coleção de documentos implica a elaboração de um modelo impreciso de representação da semântica dos textos contidos nesses documentos. A seleção de atributos é uma etapa de pré-processamento da representação dos textos que tem como objetivo eliminar atributos irrelevantes ou inapropriados. Uma das principais vantagens desse processo é reduzir o risco de *overfitting*. Outra motivação é diminuir o número de dimensões do espaço de descritores, o que pode aumentar a eficiência computacional em tempo e/ou espaço, além de simplificar o próprio modelo.

[12] é o processo utilizado para descobrir padrões interessantes e úteis em um conjunto de dados textuais.

Costuma-se então usar algumas técnicas básicas de PLN no suporte à estruturação de documentos, tais como remoção de *stopwords*[13], *stemming*[14], entre outras tarefas de normalização para controle do tamanho do vocabulário.

Outra abordagem é conhecida como *document frequency thresholding*, que elimina todos os atributos que aparecem menos do que n vezes no documento, reduzindo dramaticamente o número de atributos mesmo para valores pequenos de n. Esta abordagem está baseada na conjuntura de Apté et al. (1994), que afirma que estimativas de parâmetros para termos de baixa frequência não são confiáveis o suficiente para contribuir com informação útil.

Algumas outras técnicas estatísticas baseadas na frequência dos termos da coleção também são usadas para a redução da dimensionalidade dos atributos, a exemplo dos *Cortes de Luhn*, que delimitam o espaço de atributos pela frequência relativa de ocorrência dos mesmos na coleção textos, selecionando aqueles que são mais representativos na discriminação dos documentos. Entretanto, é percebido que nem sempre a frequência evidencia um termo, o que faz com que os sistemas baseados na estatística de frequência dos termos gerem muito "silêncio" ou muito "ruído", pois existem termos importantes (que representam conceitos) que podem aparecer com baixa ou alta frequência, respectivamente, e por isso podem ser eliminados do corpus durante a seleção dos atributos para indexação.

3.3 Dependência de termos

Muito se tem estudado sobre como capturar a textura de um texto, ou seja, seu conteúdo semântico. A própria palavra "texto" tem sua raíz vinda do Latim "texere", que significa "tecer". Tecer palavras atribuindo-lhes um sentido, sugerindo um processo de conexão entre os termos, difícil de separar. Muitos modelos de representação de textos assumem que as palavras podem aparecer aleatoriamente no texto, independentes umas das outras. A verdade é que as palavras aparecem no texto seguindo um padrão de distribuição governado pela progressão textual dos tópicos discutidos e convenções comunicativas (Katz, 1996).

[13] são palavras que refletem pouco conteúdo ou são tão comuns que não distinguem nenhum subconjunto de documentos, como as palavras de classe fechada: pronomes, conjunções, artigos.

[14] processo de conflacão de variantes morfológicas, obtendo-se a raíz, ou radical, de cada palavra.

Essa interconexão de termos é responsável pela semântica do texto, que é essencial para evitar a degradação da eficiência de sistemas de RI. Atualmente a capacidade de automatizar o entendimento do texto, ou capturar sua semântica, é limitada, muito embora existam exceções para os casos nos quais é restrito o domínio de conhecimento. Parte do problema deve ser atribuída à incompletude intrínseca dos recursos para o processamento da linguagem natural disponíveis atualmente (Baeza-Yates, 2004a).

São três as abordagens clássicas para a representação de textos: a Booleana, a vetorial e a probabilística. A abordagem Booleana é baseada na álgebra Booleana e define o critério binário de relevância. A abordagem vetorial (Salton e Lesk, 1968), com fundamentação geométrica, representa cada documento (e a consulta do usuário em um sistema de RI) como vetores de termos $\vec{d_j} = \langle w_{1j}, ..., w_{|T|j} \rangle$, onde T é o conjunto de termos encontrado em toda coleção de documentos $D = \{d_1, d_2, ..., d_{|D|}\}$. Esses vetores são usados para calcular o grau de similaridade entre documentos e consultas. A abordagem probabilística (Robertson e Sparck-Jones, 1976) aplica a Teoria da Probabilidade na RI. Outros modelos de representação de textos são encontrados na literatura (Baeza-Yates e Ribeiro-Neto, 1999), como o modelo baseado em conjuntos nebulosos (*fuzzy set model*), modelo Booleano estendido, modelo de indexação semântica latente, modelo baseado em redes neurais, redes Bayesianas, cadeias Markovianas, listas não sobrepostas, agrupamento de palavras, grafos conceituais, etc.

As estratégias que seguem essas abordagens podem ser divididas em dois grupos, quanto ao modelo adotado para representação de consulta e documentos: as que usam modelos com unigramas e as que usam modelos com dependência de termos (Gonzalez, 2005). Os modelos clássicos de representação de textos assumem que cada palavra encontrada no texto dos documentos é estatisticamente independente de todas as outras palavras, o que caracteriza o princípio da suposição da independência entre os termos (Robertson e Sparck-Jones, 1976; Robertson, 1977; Salton et al., 1982; Cooper, 1995), facilitando a formalização desses modelos e fazendo com que suas implementações se viabilizem. Dessa forma, as abordagens clássicas (Booleana, vetorial e probabilística) utilizam modelos de representação de textos baseados em unigramas, nos quais a unidade básica de indexação é representada por um e apenas um termo do documento.

O principal objetivo ao usar uma estrutura multi-termos interdependentes como unidade básica de indexação é para aumentar a precisão na captura de conceitos. Atributos

formados por múltiplos termos representam de forma mais eficiente os tópicos tratados por um documento, solucionando possíveis ambiguidades através do relacionamento entre as palavras e do papel atribuído ao contexto desse relacionamento.

Por exemplo, em modelos unigrama, documentos irrelevantes que contenham a sequência de palavras *período de execução* podem ser recuperados pela consulta *execução + fiscal*. Se a sequência de palavras isoladas e independentes uma das outras fosse substituída por unidades multi-termos, identificadas pelas suas forças de coesão frasal (motivação linguística) ou de forma estatística (*n-gram*), o documento exemplificado não seria recuperado pela respectiva consulta em um sistema de RI. O uso de conhecimento linguístico para selecionar múltiplos termos como representantes de um documento provê ganhos de precisão, por permitir distinções mais consistentes entre termos similares (desambiguação) porém não idênticos, com estrutura interna diferente, ou estabelecendo relações mais elaboradas entre os termos.

É percebido um número crescente de trabalhos que adotam modelos com dependência de termos (Evans e Zhai, 1996), ou termos de índice baseados em expressões. Esses estão fundamentados na inexistência ou no relaxamento da suposição de independência dos termos, devido às inconsistências nela encontradas e às regularidades que podem ser consideradas em modelos com dependência de termos (Gonzalez, 2005).

Os modelos com dependência de termos podem ser fundamentados por um conjunto de relacionamentos estatísticos entre os termos (motivação estatística), regido através da probabilidade de co-ocorrência contígua (*n-gram*) entre eles (Miller et al., 1999; Caropreso et al., 2001). Muitos outros trabalhos estão fundamentados em conhecimento linguístico (Liu et al., 2004; Lee e Lee, 2005; Vilares et al., 2002) (motivação sintática), situando-se expressivamente aqueles relacionados aos *sintagmas nominais*[15], representando a menor parte do discurso portadora de informação. Os múltiplos termos cujo uso são mais defendidos na RI são os constituídos por sintagmas nominais (Arampatzis et al., 2000b).

Alguns experimentos (Apté et al., 1994; Sahami, 1998) demonstraram que representações mais sofisticadas envolvendo dependência de termos, às vezes, não agregam significativamente melhor desempenho em relação aos modelos unigrama. A razão mais provável

[15] são os constituintes imediatos de uma sentença com comportamento sintático de sujeito, de objeto direto ou indireto, ou de adjunto adnominal, se precedido de preposição (Perini, 2000)

para explicar esses resultados é que, embora unidades de índice mais sofisticadas tenham qualidade semântica superior, a qualidade estatística é inferior em relação a termos baseados em palavras simples (Lewis, 1992). No campo do AM, de acordo com Joachims (2002), a abordagem unigrama (também conhecida como *bag-of-words*) confere uma boa relação entre expressividade e complexidade. Enquanto representações mais expressivas capturam melhor o significado do documento, sua complexidade é maior e degrada a qualidade de modelos estatísticos.

Modelos de representação de textos com dependência de termos foram experimentados na recuperação de documentos nas conferências TREC [16]. A principal conclusão sobre a inviabilidade do uso de técnicas de PLN em cenários reais de RI foi devido ao alto custo computacional demandado pelos seus algoritmos frente ao pequeno aumento de eficiência produzido. Basicamente o pequeno aumento deve-se aos erros do PLN na deteção de estruturas complexas e ao uso dessas estruturas como descritores tão informativos quanto àquelas usadas no modelo unigrama. Segundo Arampatzis et al. (2000b), as técnicas de PLN disponíveis no final do século XX eram desprovidas de eficiência e acurácia suficientes para quebrar o paradigma *bag-of-words* com resultados mais consistentes e não duvidosos. Adicionalmente a este fato, existe a falta de acurácia das atuais ferramentas disponíveis para desambiguação de conceitos [17]. Em contraste, quando a desambiguação é feita manualmente, resultados promissores sobre a mesma atividade foram obtidos (Voorhees, 1993).

Muito embora existam evidências nesse sentido, a palavra final sobre o uso de modelos com dependência de termos ainda não foi deferida, e pesquisas nessa direção ainda estão sendo ativamente conduzidas (Sebastiani, 2002), algumas delas com resultados a favor do modelo com dependência de termos. De acordo com Sebastiani (2002), a combinação das duas abordagens (estatística e linguística) é, provavelmente, o melhor caminho a seguir. Em Arampatzis et al. (2000a), onde é analisado esquemas de indexação linguisticamente motivados, conclui-se que o PLN e outros recursos linguísticos serão partes indispensáveis de todo sistema eficiente de RI.

Esse trabalho pretende seguir a linha de pesquisa proposta por um modelo híbrido de representação de documentos com dependência de termos, baseado na combinação de sintagmas nominais (conhecimento linguístico) juntamente com a co-ocorrência de unigra-

[16] Text REtrieval Conference - http://trec.nisc.gov
[17] WSD - Word Sense Disambiguation tools

mas dentro dos sintagmas mais representativos da coleção de documentos (conhecimento estatístico).

3.4 Representatividade dos atributos

Adicionalmente ao critério de seleção das unidades de indexação, encontra-se um outro fator que, juntos, determinam a escolha entre os diversos modelos de representação de textos existentes: o cálculo da representatividade dessas unidades. Os termos de índice devem refletir valores que quantificam sua representatividade na coleção de documentos, ou seja, estima-se o quanto os atributos são representativos do texto, ou ainda, determina-se a relevância do conceito descrito por essas unidades de indexação. Conforme demonstrado na Tabela 1, o grau de representatividade de um descritor t em um documento d é dado pelo peso $w_{t,d}$.

A mais simples representação desse valor é a Booleana, que indica a presença ou ausência do termo no documento em questão, através dos valores *0* ou *1*, respectivamente. Todas as operações computacionais envolvidas nesse modelo atendem às condições impostas pela algebra de Boole. Na representação binária, o valor w_{ij} é igual a 1 se o termo t_j ocorre no documento d_i e igual a 0 caso contrário, para $j \in \{1, \ldots, M\}$ e $i \in \{1, \ldots, N\}$, conforme a Equação 2.

$$w_{ij} = \begin{cases} 1, & \text{se } t_j \in d_i \\ 0, & \text{caso contrário.} \end{cases} \tag{2}$$

Apesar de simples, são bem conhecidas as desvantagens do modelo Booleano em sistemas de RI (Baeza-Yates e Ribeiro-Neto, 1999): (i) sua estratégia de operação é baseada no critério de decisão binária (sim ou não) e na teoria dos conjuntos, não permitindo gradientes de intervalos, o que degrada a performance dessas operações; (ii) o processo de tradução de uma requisição em uma expressão Booleana não é simples para a grande maioria dos usuários de sistemas de RI; (iii) o modelo Booleano traz um grande volume de documentos irrelevantes, assemelhando-se mais a um modelo de recuperação de dados do que propriamente de informações. Analogamente, em atividades relacionadas ao campo de

AM, a exemplo de Categorização de Textos (CT), o modelo Booleano é, por muitas vezes, inadequado.

A medida mais comumente usada em sistemas de RI e AM para a representatividade dos descritores é denominada *tf.idf* [18] (Salton e Buckley, 1987b) que, além de acusar a presença do termo, reflete informações sobre a sua frequência no documento multiplicada pelo inverso da frequência desse termo na coleção.

A medida tf (*term frequency*) utiliza o número de ocorrências de t_j em d_i. A idéia é que termos mais frequentes no documento sejam mais relevantes que aqueles menos frequentes. Nesse caso, w_{ij} assume o valor $tf(t_j, d_i)$, que representa o número de vezes que o termo t_j ocorre no documento d_i, conforme a Equação 3.

$$w_{i,j} = tf(t_j, d_i) \tag{3}$$

Todavia, existem termos muito frequentes no conjunto de documentos e, portanto, não carregam uma "força" de representação significativa na discriminação do seu conteúdo. O fator de ponderação *idf* (*inverted document frequency*) garante que termos muito comuns sejam penalizados, favorecendo aqueles que aparecem em poucos documentos, conforme a Equação 4.

$$idf_t = \log \frac{|D|}{df_t} \tag{4}$$

Assim, as medidas tf e idf podem ser combinadas em uma nova medida denominada $tf.idf$. O valor de w_{ij} pode então ser calculado conforme a Equação 5.

$$w_{i,j} = tfidf(t_j, d_i) = tf(t_j, d_i) \times \log \frac{|D|}{df_t} \tag{5}$$

É comum normalizar a medida tf a fim de não beneficiar injustamente os documentos longos. Usa-se dividir tf pelo número de termos do documento ou pelo número de

[18] Term Frequency x Inverted Document Frequency

ocorrências do termo mais frequênte no documento. Apresentaremos no Capítulo 6 um contexto mais apropriado para referenciar a normalização do componente tf.

No modelo probabilístico, o cálculo do peso de um descritor t em um documento d corresponde à formula Okapi BM25 (Robertson e Walker, 1994), descrito pela Equação 6:

$$W_{t,d} = \frac{w_{t,d}(k_1 + 1)}{k_1((1 - b) + b\frac{DL_d}{AVDL}) + w_{t,d}} \times idf_t \tag{6}$$

onde:

- k_1 é um parâmetro para correção de frequência; usualmente assume o valor 1,2;
- b é um parâmetro para controle das hipóteses do escopo[19] e da verbosidade[20]; usualmente assume o valor 0,75;
- DL_d é o comprimento (quantidade de caracteres ou palavras) do documento d;
- $AVDL$ é o comprimento médio dos documentos da coleção.

Uma alternativa para o modelo algébrico (vetorial) é a indexação semântica latente, cuja idéia principal é mapear cada documento (inclusive o vetor de consulta) para uma matriz de menor dimensionalidade, associada a conceitos. Isto é devido à alta esparsividade do modelo vetorial, que pode conduzir a uma baixa performace de recuperação, seja por não retornar documentos relevantes ou por retornar documentos irrelevantes para uma determinada consulta. Os atributos do novo espaço vetorial representam, de alguma forma, significados/sentidos dos termos, que agora encontram-se agrupados segundo algum critério relacional.

3.5 Evidência dos descritores

Evidência é a condição do que se destaca, é a qualidade do que é evidente e, por sua vez, evidente é aquilo que não oferece ou não dá margem à dúvida (Ferreira, 1999; Houaiss, 2002).

[19] documentos mais longos têm mais informação que os menos longos

[20] documentos mais longos possuem escopo similar ao de um documento menos longo, simplesmente usam mais palavras

O modelo de representação de documentos com dependência de termos proposto neste trabalho utiliza o conceito de evidência para compor o cálculo da representação de seus descritores, em adição à frequência de ocorrências dos mesmos. Quanto maior a evidência do descritor (destaque sem ambiguidade), maior sua representatividade.

A abordagem probabilística é adotada para o cálculo do peso dos descritores, através de uma adaptação da equação Okapi BM25, conforme a Equação 7:

$$W_{t,d} = \frac{w_{t,d}(k_1 + 1)}{k_1((1 - b) + b\frac{DL_d}{AVDL}) + w_{t,d}} \tag{7}$$

Nesta equação, $w_{t,d}$ é a evidência do descritor t no documento d, calculada da seguinte forma:

$$w_{t,d} = k_2.f_{t,d} + \sum_{s} f_{s,t,d} \tag{8}$$

onde:

- k_2 é um coeficiente de amortização de frequência, usualmente 0,5;
- $f_{t,d}$ é a frequência de ocorrência de t em d e;
- $f_{s,t,d}$ é a quantidade de SNs s em d, onde $t \in s$.

e para um SN s, a evidência em um documento d, representada por $w_{s,d}$, é:

$$w_{s,d} = k_3.f_{s,d} \times \sum_{i=1}^{n} w_{t_i,d} \tag{9}$$

onde:

- k_3 é um coeficiente de amortização de frequência, inicialmente 1;
- $f_{s,d}$ é a frequência de ocorrência de s em d e;

3.6 Sintagmas nominais evidentes

Sintagmas nominais (SN) constituem um caso especial de relacionamento multi-termos porque carregam informações com alto poder discriminatório e potencial informativo. Segundo Kuramoto (2002), essas estruturas estabelecem referências a um conceito, objeto ou fato do mundo real. O sintagma é um conjunto de elementos que constituem uma unidade significativa numa sentença e que mantêm entre si relações de dependência e de ordem (Lobato, 1986).

Antigamente o processo de indexação de documentos era artesanal, no qual técnicos especializados eram munidos de vocabulários controlados, tesauros, tabelas e listas que forneciam os descritores adequados à elaboração da representação de cada documento. De forma análoga, os elementos que devem ser extraídos de um documento para representá-lo devem possuir a mesma função de um descritor (Kuramoto, 2002), sendo que a automatização desse processo preserva seu conceito e natureza funcional. O índice então formado por esse processo com seus respectivos graus de representatividade (pesos) compõem um espaço de descritores.

A identificação de SNs em textos tem aplicações em diversos problemas: recuperação e extração de informações, análise sintática, resolução de co-referência, identificação de relações semânticas, entre outros. Nosso foco é compor um modelo híbrido de espaço de descritores, composto por SNs e termos unigrama, cuja representatividade individual de cada descritor é fortemente influenciada pela sua evidência no texto.

O uso de programas computacionais para o reconhecimento e extração de componentes sintáticos no texto, a exemplo dos sintagmas nominais, caracterizam-se por dois tipos de conhecimento: (i) os de motivação simbólica e (ii) os de motivação estatística. Alguns trabalhos (Voutilainen, 1993; Miorelli, 2001) usaram informação simbólica no reconhecimento de SNs, através da aplicação de regras gramaticais criadas e mantidas manualmente, o que é um processo custoso em relação aos atuais métodos estatísticos e, muitas vezes, não é reaproveitável por diferentes aplicações (Santos, 2005).

Uma máquina de estado finito não determinística (*NFA* [21]) é, convencionalmente, o modelo de representação mais simplista para a descrição do processo simbólico, conforme

[21] Non-deterministic finite automata

ilustrado na Figura 1[22], para a língua inglesa. A língua portuguesa e inglesa guardam similaridades no nível sintático [23]. Uma das principais diferenças em relação à estrutura do SN nessas línguas é a posição do adjetivo, preferencialmente pré-nominal no inglês e pós-nominal no português.

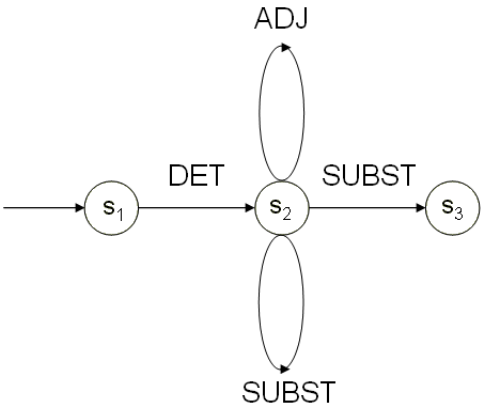


Figura 1: NFA para reconhecimento de SNs

Técnicas estatísticas de AM têm se revelado uma potente ferramenta na viabilização de tarefas linguísticas, a exemplo de etiquetagem morfossintática, identificação de sintagmas nominais, correção ortográfica, reconhecimento de entidades nomeadas, identificação dos limites das sentenças, etc.

A identificação dos sintagmas nominais através de AM tem sido pesquisada intensivamente na última década (Church, 1988; Brill, 1995; Ramshaw e Marcus, 1995; Kudo e Matsumoto, 2001). Classificada como uma atividade reconhecidamente difícil [24], sua extração às vezes limita-se apenas a certos tipos de sintagmas, se forem empregados apenas métodos determinísticos para sua identificação. Existem recentes publicações, inclusive, que exploram alguma metodologia específica para essa atividade em um domínio particular, utilizando AM, a exemplo da área biomédica (Wermter et al., 2005).

No Brasil, alguns trabalhos específicos para a língua portuguesa foram conduzidos para a identificação de SNs básicos [25] (Santos, 2005), bem como aquelas estruturas recursivas contendo pós-modificadores como adjetivos e sintagmas preposicionados. Como

[22] extraída do livro (Jackson e Moulinier, 2002), pág. 91
[23] ordem das palavras com relação às classes gramaticais
[24] Handbook of Computational Liguistics, Ralph Grishman, capítulo 30: Information Extraction
[25] são estruturas não recursivas que incluem determinantes e modificadores

resultado, o problema de identificar SNs do português torna-se mais complexo do que a identificação apenas de SNs básicos (como acontece com os da língua inglesa), visto que inclui o problema da ligação do sintagma preposicionado.

Várias técnias foram aplicadas para a escolha das sequências de SNs com destaque para SVM (*Support Vector Machines* (Cortes e Vapnik, 1995)) e TBL (*Transformation Based Learning* (Brill, 1995)). Segundo pesquisa realizada com TBL (Ngai e Yarowsky, 2000), é mais vantajoso utilizar recursos humanos para fazer anotação do corpus e utilizá-la para treinar um identificador de SNs do que para criar e manter manualmente um conjunto de regras para uma gramática de identificação de SNs. Entretanto, utilizamos em nosso experimento conhecimento linguístico e estatístico, em função da disponibilidade das ferramentas computacionais encontradas para o caso do português brasileiro, cujo desenvolvimento não é foco da presente proposta, apesar desta ser altamente dependente de seus resultados.

Apresentaremos no próximo Capítulo o processo de Filtragem de Informação (FI), diferenciando-o da Recuperação de Informação pela natureza temporal das necessidades de busca do usuário e pelo dinamismo dos dados acessados. É explicado o processo exploratório praticado pelo usuário em ciclos de iterações com o sistema, bem como a viabilidade de uso de Aprendizado de Máquina (AM) para potencializar o processo de FI, ao apresentar os filtros como naturais categorizadores automáticos de textos. Por fim, introduzimos as principais famílias de algorítmos de AM utilizados nesse domínio.

4 Indução automática de filtros

Ao esclarecer a diferença entre os processos de filtragem e recuperação de informação, pretendemos conduzir o leitor a perceber o quanto esses conceitos são complementares para a elaboração de uma estratégia mais eficiente de acesso à informações relevantes a um determinado tópico, enriquecendo a experiência final do usuário. Não obstante, almejamos compreender porque o processo de generalização a partir de exemplos, denominado inferência indutiva, apresenta-se como uma alternativa viável para o processo de FI, ao observar que sua natureza é essencialmente classificatória.

4.1 Filtragem de Informação

A Internet tornou-se o veículo mais eficiente na disseminação de conhecimento em tempo quase real, provendo um volume dinâmico de dados multimídia e novas formas de interação que bem caracterizam a sociedade da informação. Entretanto, esse fenômeno acentuou expressivamente o problema do tratamento de informações irrelevantes, erradas, obsoletas ou mesmo indesejáveis pelos seus usuários.

Conviver com a sobrecarga de informação gerada pela Internet como hoje a conhecemos está se tornando um problema cada vez mais crítico e esse é o papel principal da FI. Entre seus objetivos, um dos mais importantes é a ampliação qualitativa da experiência do usuário no processo exploratório de busca por informação relevante, através da seleção analítica em um fluxo contínuo de dados.

O processo de Filtragem de Informação atua como mediador entre as fontes de informação e seus usuários finais, controlando seletivamente a distribuição de informação (Sheth, 1994). A FI é considerada um dispositivo que poupa tempo e esforço dos seus usuários (Baclace, 1991) ao priorizar quais informações poderão tomar sua atenção segundo algum critério de relevância.

Duas abordagens especiais de acesso à informação consolidaram-se ao longo da história como estratégias analíticas de busca: a Recuperação de Informação e a Filtragem de Informação. Sistemas de RI objetivam atender ao usuário com necessidade momentânea

de informação, cujo foco de interesse é bastante dinâmico. Supõe-se que o banco de dados consultado seja relativamente estático em sistemas de RI. Em contrapartida, sistemas de FI são projetados para prover ao usuário acesso às fontes de informações altamente dinâmicas (como em um fluxo contínuo de dados), assumindo que seus interesses sejam relativamente estáticos sobre o tempo.

O termo Filtragem de Informação descreve uma variedade de processos envolvendo a entrega ou disseminação de informação para aqueles que precisam dela. Segundo Belkin e Croft (1992), a FI tem sido usada para descrever o processo de distribuição e roteamento de informações advindas de bancos de dados remotos, através do qual as informações selecionadas são resultados das requisições (acesso e recuperação) de buscas neles efetivados. Assim, a FI constitui um processo de extração/seleção de porções relevantes ou úteis de grandes repositórios de dados ou de fluxos contínuos de textos baseados em padrões relativamente estáticos de interesse dos seus usuários.

Nesse cenário, a principal diferença entre um engenho de busca e um sistema de FI é que, enquanto o primeiro acha documentos relevantes em uma base de dados relativamente estática, o segundo seleciona ou remove aqueles documentos irrelevantes advindos do fluxo dinâmico de documentos retornado pelo engenho de busca. Os usuários enxergam apenas os documentos que foram extraídos (filtrados) do fluxo. São sistemas complementares que potencializam a experiência do usuário final de um sistema de RI.

A natureza do interesse do usuário é responsável por definir como as duas estratégias irão se combinar para endereçar o problema do acesso à informação de forma mais apropriada. Por exemplo, as preferências do usuário por gênero de música ou filme certamente são mais estáveis ao longo do tempo em que houver necessidade de acesso por esse tipo de informação. Certas aplicações têm uma taxa de mudança das fontes de informação (como no fluxo dinâmico de dados) maiores do que outras, bem como existem aplicações que possuem uma taxa de mudança de necessidades de informação diferentes de outras (Baudisch, 2001). O desafio de encontrar o *trade-off* entre a aplicação de técnicas híbridas de RI e FI para uma determinada aplicação é ilustrado na Figura 2.

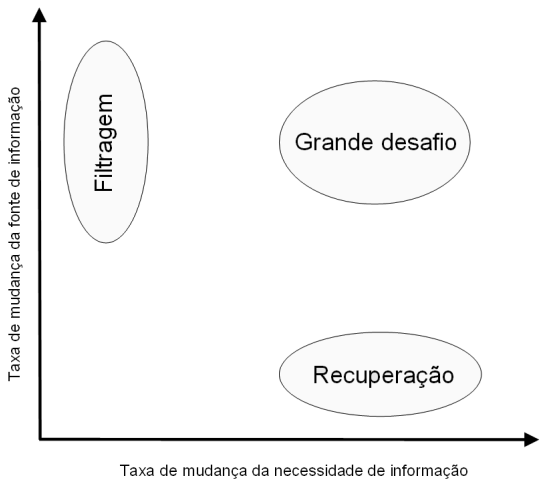


Figura 2: Desafio da aplicação no emprego de técnicas de RI e FI

4.2 Perfil de busca adaptativo

Sistemas personalizados de FI incorporam interesses do usuário ou de um grupo de usuários, expressos através de um *perfil de busca* (usualmente representado através de uma tabela atributo-valor[26], como ilustrado na Tabela 2), que consolida a representação dos textos considerados relevantes sobre um domínio (ou um conjunto de domínios) de interesse do usuário.

	t_1	t_2	$\ldots$	$t_{\lvert T\rvert}$	C
d_1	w_{11}	w_{12}	$\ldots$	$w_{1\lvert T\rvert}$	c_1
d_2	w_{21}	w_{22}	$\ldots$	$w_{2\lvert T\rvert}$	c_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$d_{\lvert D\rvert}$	$w_{\lvert D\rvert 1}$	$w_{\lvert D\rvert 2}$	$\ldots$	$w_{\lvert D\rvert\lvert T\rvert}$	$c_{\lvert C\rvert}$

$$\textit{onde} \begin{cases} d_1 \text{ a } d_{|D|} \text{ são os documentos da coleção;} \\ t_1 \text{ a } t_{|T|} \text{ compõem o espaço de descritores;} \\ c_1 \text{ a } c_{|C|} \text{ são as categorias predefinidas, sendo } |C| \leq |D| \text{ e;} \\ w_{11} \text{ a } w_{|D||T|} \text{ são os pesos relacionados a cada descritor.} \end{cases} \tag{10}$$

Tabela 2: Tabela atributo-valor com categorias predefinidas

Essas necessidades de informação caracterizam uma *função de relevância*, que consiste em um mapeamento entre um espaço de objetos quaisquer e um espaço de objetos realmente

[26] representação estruturada de textos na qual consta a associação entre cada um dos seus atributos e o respectivo valor (peso) de sua representatividade

relevantes para o usuário. Esse mapeamento fundamenta o problema da representação dos interesses do usuário em relação à relevância que cada objeto desse espaço tem para ele (Lam et al., 1996).

O termo *Perfil* nasceu com o trabalho de Luhn (1958), no qual foi concebido um sistema automático para disseminação de informações às várias seções de qualquer organização. Esse sistema "inteligente" era capaz de usar "perfis de interesse" criados manualmente por bibliotecários, utilizados pelo processo de seleção de textos para cada seção da organização. Ao descrever a função do módulo de seleção como "disseminação seletiva de nova informação", ele cunhou o termo que descreveu este campo por quase meio século (Oard e Marchionini, 1996).

Em analogia aos sistemas de RI, o perfil de busca do usuário de um sistema de FI pode ser considerado como uma consulta (*query*) estável durante um certo intervalo de tempo, que retorna uma vasta gama de objetos relacionados aos seus interesses persistentes. Os perfis são caracterizados por serem mais duradouros e consistentes que as consultas de um sistema de RI, que tipicamente encontram-se relacionadas com necessidades de informação específicas e momentâneas.

Enquanto a RI trata problemas inerentes à adequação da consulta como uma representação aderente à necessidade de informação momentânea do usuário, a FI já pressupõe que o seu perfil de busca seja uma especificação correta dos seus interesses mais duradouros (Belkin e Croft, 1992).

Por mais duradouro que sejam seus perfis, para que um sistema de FI seja realmente útil, ele precisa permitir que os mesmos se adaptem às novas necessidades de informação dos usuários. A capacidade de aprendizado do perfil de busca pode adaptá-lo às mudanças de preferências e interesses do seu usuário, bem como automatizar algumas de suas principais tarefas, refletindo o seu potencial evolutivo. Aprendizado e adaptação são tópicos amplamente abordados nas áreas de RI e FI no treinamento do sistema para melhor representar e trabalhar as necessidades dos seus usuários.

Esse processo acontece durante todo o tempo em que o usuário vivencia a experiência de busca por novas informações, observando e aprendendo com seus interesses e hábitos. O aprendizado adaptativo é plausível em ambientes onde é possível a observação contínua

do fluxo de dados e das mudanças de estados do sistema, reflexo das decisões tomadas pelo usuário.

A proposta desse trabalho não leva em consideração a dinâmica da mudança de foco do usuário ao longo de sua experiência de busca por informações, muito embora isso seja possível através da customização dos algoritmos aqui utilizados para o processo de FI. Entretanto, seria producente considerar, além dos algoritmos abordados, uma arquitetura modular do perfil de busca adaptativo como sugerido por Baudisch (2001). Procuramos nos concentrar apenas no impacto que o modelo linguisticamente motivado (envolvendo os sintagmas nominais evidentes no texto) provoca no processo de FI, e como esse impacto reflete nas métricas de avaliação de um sistema de RI.

4.3 O processo iterativo de aprendizado

Em um sistema personalizado de FI, a interação entre o sistema e o usuário se dá em ciclos. Assumindo que as ações do usuário sejam consistentes, durante o processo exploratório espera-se uma melhoria gradativa dos resultados do sistema, advindo do aprendizado incremental proporcionado pelo ciclo iterativo descrito a seguir:

Quando o usuário acessa o sistema pela primeira vez ele o alimenta com textos que melhor representam o seu campo inicial de interesse. O sistema então formaliza essa ação através da construção de um perfil baseado nas informações adquiridas. O perfil assemelha-se a uma tabela atributo-valor como ilustrada na Tabela 2.

Ao receber um fluxo de documentos advindo de uma consulta a um sistema de RI remoto, o sistema coadjuvante de FI estrutura cada um deles da mesma forma que o perfil inicial, refletindo o mesmo modelo de representação de textos para ambas entidades (perfil e documento retornado).

O filtro entra em ação e é acionada uma análise comparativa entre cada documento e o perfil de busca. O resultado desta análise é binário, podendo cada documento ser classificado apenas como relevante ou irrelevante em relação ao perfil. São apresentados ao usuário apenas os documentos relevantes.

Os documentos relevantes listados pela interface do sistema local de FI trazem consigo um grau de confiabilidade, que reflete uma medida de semelhança com o perfil do usuário. Muito provavelmente baseado nesse valor é que o usuário decidirá quais documentos irá acessar para leitura.

Para cada um dos documentos lidos, o usuário devolve ao sistema uma indicação positiva, negativa ou, na ausência de indicação, o sistema presume que essa seja nula e, portanto, desconsiderada. Esta ação é conhecida como realimentação explícita, pois confere a participação consciente do usuário ao fornecer informação ao sistema. Por outro lado, o sistema também pode observar algumas evidências implícitas sobre os interesses do usuário em cada documento visitado. Uma combinação de eventos observados pode ter relação com a relevância de um determinado documento. Por exemplo, o tempo gasto pelo usuário lendo o documento em razão do seu tamanho (Ngu e Wu, 1997), o próprio ato de salvar o documento, etc.

A realimentação é registrada para cada documento apresentado e usada para melhorar a performance do sistema de FI através da adaptação do perfil inicial do usuário. O ciclo recomeça na próxima interação entre o usuário e o sistema de RI que, inclusive, pode não ser o mesmo usado na iteração anterior. As iterações também não precisam ser executadas em momentos imediatamente seguintes umas das outras, visto que o perfil é uma estrutura persistente, que evolui com a experiência do usuário.

Com esse processo é conferida mais uma vantagem na utilização de um sistema de FI com papel coadjuvante aos sistemas de RI no aprimoramento qualitativo da experiência do usuário: é desnecessário tê-los juntos em um mesmo ambiente físico. Assim, o sistema de FI pode ser instalado e personalizado no ambiente local do usuário, enquanto que os sistemas de RI são consultados remotamente.

A performance e modularidade conferidas ao sistema de FI o elege como uma ferramenta promissora no processo de RI, à medida que as técnicas híbridas que utilizam conhecimento estatístico e linguístico envolvidas na representação e categorização de textos evoluam à luz dos benefícios trazidos para seus usuários.

4.4 Filtros colaborativos e cognitivos

Atualmente os sistemas de FI são classificados como sociais ou colaborativos e cognitivos ou baseados em conteúdo.

Na filtragem colaborativa (ou social), os documentos são selecionados com base nas anotações feitas por um conjunto de usuários que compartilham interesses comuns. São amplamente utilizados por sistemas de recomendação (Goldberg et al., 1992), fazendo uso de um banco de preferências dos usuários cadastrados. Ali encontram-se aqueles que compartilham os mesmos interesses, sendo possível predizer quando um item de informação desconhecido é potencialmente interessante para um determinado usuário, baseando-se em como os outros classificaram esse item em função das suas próprias experiências.

Um exemplo expressivo de sucesso de filtros colaborativos são os *listmanias* da Amazon[27], onde são indicados aos consumidores itens de consumo potencialmente relevantes, baseando-se na experiência de navegação entre os itens oferecidos e recomendações de outros consumidores que já tiveram experiência similar.

Os filtros cognitivos, também conhecidos como filtros baseados em conteúdo, como o próprio nome indica, fazem uso dos componentes e estruturas latentes do texto para determinar a relevância do documento em função do perfil de busca, que constitui o próprio filtro em questão.

4.5 Filtros como categorizadores de textos

O processo seletivo de filtrar os documentos relevantes dos irrelevantes é classificatório, no sentido que existe um critério de classificação para separar dois conjuntos mutuamente exclusivos de documentos. Não importa, nesse momento, se esse critério usa o conteúdo (texto) do documento ou quaisquer outros de seus atributos para a classificação. Poderíamos supor, por exemplo, que documentos relevantes, para um certo contexto, são aqueles cuja data de última atualização pertença ao ano corrente.

[27] www.amazon.com

Geralmente os documentos carregam consigo dados ou metadados, como sua fonte de edição, título, data, autor, palavras-chave que melhor os qualificam, etc. Essas informações, denominadas *exógenas*, podem servir para classificar[28] os documentos, em função de algum critério estabelecido entre elas. Todavia, estaremos voltados para as técnicas de categorização[29] de documentos baseado em seu conteúdo, ou seja, nas informações *endógenas* contidas no texto, compreendendo estruturas sintáticas, relacionamento entre termos, semântica, estilo e outros tantos componentes que influenciam a desambiguação de sua identidade. Na Figura 3[30], é ilustrada uma visão geral sobre o processo de classificação de documentos, onde d_1 a d_n são os documentos e C_1 a C_k são as categorias.

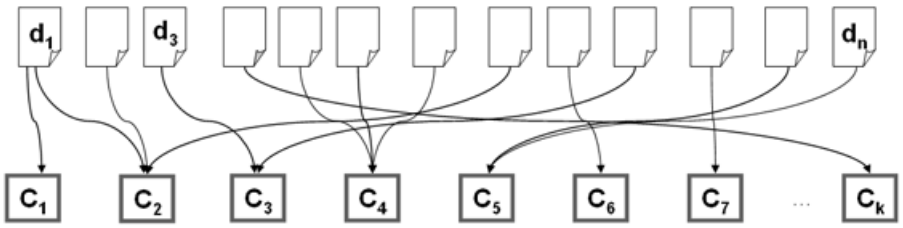


Figura 3: Classificação de documentos

Torna-se evidente que o filtro cognitivo pode ser descrito como um categorizador binário, pelo fato de que os documentos resultantes do processo de categorização ou são relevantes ou irrelevantes em relação ao classificador. As duas únicas categorias existentes são mutualmente exclusivas ou disjuntas.

A classificação automática de textos começou a ser estudada na década de 60, mas somente tornou-se viável com o avanço de *hardware* e *software*. Durante a década de 80, esse processo era realizado através da criação manual de regras de composição de textos, processo que envolvia o conhecimento de especialistas na área de discurso que abrange os conceitos a serem descritos nas categorias. Então, os primeiros métodos para a automatização da classificação de textos eram baseados na manufatura de regras através de conhecimento especializado sobre um determinado domínio. Essas regras servem para compor o critério de categorização fundamentado no reconhecimento de padrões entre cadeias de caracteres, geralmente através de máquinas de estados finitos e *parsers* poderosos, como aqueles que permitem encontrar simultaneamente múltiplos padrões de textos por simila-

[28] termo que abrange qualquer tipo de associação entre documentos e classes
[29] termo mais restrito usado para associar documentos apenas em função do seu conteúdo
[30] extraída de www.umiacs.umd.edu/ joseph/tex-categorization.ppt, pág. 5

ridade (Navarro et al., 2003). A precisão e a revocação de tal processo são extremamente dependentes do domínio da aplicação e da eficácia na elicitação do conhecimento do especialista e no seu respectivo processo de representação no ambiente computacional. Essa representação estruturada do conhecimento especializado servirá como função de classificação para o categorizador, nesse caso um autômato de estados finitos. É um processo custoso e que envolve muito esforço humano durante a fase de elicitação do conhecimento e sua representação, além de ser pouco flexível em relação às mudanças a que esses padrões estão sujeitos.

Somente a partir dos anos 90 começou a ser utilizado o paradigma de aprendizagem computacional para categorização de textos. Esses métodos têm tido evidência no meio acadêmico e mais recentemente na indústria. A nova abordagem está fundamentada no campo do Aprendizado de Máquina (AM), que constitui uma sub-área da IA que estuda métodos computacionais relacionados à aquisição de novos conhecimentos (Mitchell, 1997). Através do AM, um processo indutivo contrói automaticamente um classificador de textos, por meio do aprendizado por exemplos previamente classificados. Quando isso acontece, é dito que o aprendizado é *supervisionado*, porque são fornecidos exemplos positivos e negativos para o treinamento do classificador para cada categoria envolvida no processo.

Formalmente, Categorização de Textos (CT) é a atividade de relacionar um valor Booleano para cada par $\langle d_j, c_i \rangle \in D \times C$, onde D é o domínio de documentos, e $C = \{c_1, ..., c_{|C|}\}$ é o conjunto de categorias predefinidas. O valor V associado a $\langle d_j, c_i \rangle$ indica que o documento $d_j \in c_i$, enquanto que o valor F associado a $\langle d_j, c_i \rangle$ indica que $d_j \notin c_i$. Sendo assim, um categorizador é uma função $\Phi : D \times C \rightarrow \{V, F\}$, denominada hipótese ou modelo, que descreve como os documentos deveriam ser classificados. A Figura 4[31] exemplifica esta função de mapeamento.

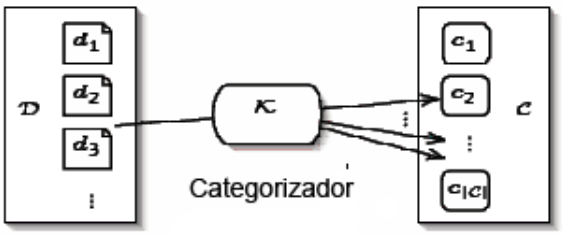


Figura 4: Função de mapeamento entre documentos e categorias

[31] extraída de www.umiacs.umd.edu/ joseph/tex-categorization.ppt, pág. 6

Segundo Sebastiani (2002), as vantagens dessa abordagem são a acurácia comparada àquelas conquistada por especialistas humanos, e o considerável ganho em termos de poder de trabalho especializado, uma vez que não há a intervenção de engenheiros do conhecimento ou de especialistas no domínio para a construção do classificador. Outras vantagens como a portabilidade para diferentes tipos de categorias ou aplicações e a flexibilidade do processo de aprendizado para adaptar-se a novas situações fazem com que essa abordagem sobrevaleça às outras.

A capacidade de aprender e adaptar-se a novas situações são essenciais para um comportamento inteligente, e um dos principais objetivos da IA é a automatização de processos nos quais, até o momento, o ser humano ainda tem um melhor desempenho (Rich, 1983). Assim é o processo de filtragem cognitiva, uma aplicação da categorização de textos.

4.6 Indução de filtros linguísticos

Indução é a forma de inferência lógica que permite que conclusões gerais sejam obtidas de exemplos particulares. A inferência indutiva constitui um dos principais meios de adquirir novos conhecimentos (principal desafio do AM) e prever eventos futuros. O aprendizado indutivo é o processo de inferência indutiva realizada sobre fatos, situações ou casos observados, os quais são fornecidos ao aprendiz por um professor ou oráculo (Mitchell, 1997). O aprendizado indutivo através de exemplos tem a tarefa de induzir descrições gerais de conceitos, utilizando exemplos específicos desses conceitos (Carbonell et al., 1984).

O modelo preditivo de indução do perfil de busca e estruturação dos documentos, bem como a respectiva atividade de CT empregada nesse trabalho, são fundamentados no campo de AM, que descreve um processo através do qual um sistema melhora seu desempenho. Segundo Simon (1983), aprendizado denota mudanças no sistema, que são adaptáveis no sentido em que elas possibilitam que o sistema faça a mesma tarefa (ou tarefas) sobre uma mesma população, de uma maneira mais eficiente a cada vez.

O processo indutivo permite que os padrões estatísticos e linguísticos do texto sejam usados para modelar a distribuição dos seus descritores juntamente com seus respectivos pesos de representatividade entre os documentos. Essa abordagem tem sido amplamente

adotada em vários trabalhos recentes de categorização de textos (Sebastiani, 2002). Existe uma forte tendência constatada atualmente para a construção do perfil do usuário fundamentada em técnicas de AM, não apenas pelo avanço dos resultados obtidos no processo eficiente de categorização de documentos, como também pela flexibilidade de seus modelos em se adaptar e evoluir através de suscessívos aprendizados sobre o domínio explorado, obtidos a partir do comportamento do usuário.

Tendo em vista que a categorização indutiva tem como base o conhecimento endógeno baseado apenas na sua semântica, e dado que a semântica de um documento é um conceito *subjetivo*, conclui-se que a pertinência de um documento à uma classe não pode ser decidida deterministicamente (Sebastiani, 2002). Ou seja, duas pessoas diferentes podem classificar o mesmo documento em categorias diferentes, ou simultaneamente em múltiplas categorias, dependendo do julgamento subjetivo desses especialistas.

Consequentemente, a automatização do processo de inferência indutiva na categorização automática de textos tem beneficiado o campo da Filtragem de Informação (Amati, Crestani, Ubaldini e De Nardis, 1997; Schapire, Singer e Singhal, 1998; Iyer, Lewis, Schapire, Singer e Singhal, 2000; Tauritz, Kok e Sprinkhuizen-Kuyper, 2000).

Nesse trabalho é introduzido o conceito de Filtro Linguisticamente Motivado (FLM), que descreve um filtro cognitivo induzido através de técnicas de AM, cujos descritores são estruturados através de uma abordagem híbrida (unigrama e com dependência de termos) baseada em conhecimento estatístico e linguístico, através do uso de sintagmas nominais (SNs), nos quais os pesos de representatividade são calculados através da evidência explícita desses descritores no texto.

O conhecimento linguístico utilizado no processo de FI acontece no momento da construção do espaço de descritores que representa a coleção trabalhada, exercendo influência sobre o quanto os algoritmos de AM conseguem generalizar. Apresentaremos no Capítulo 6 um experimento de FI sobre um contexto específico, onde o espaço de descritores é formado por SNs extraídos dos textos, juntamente com outros que são formados a partir de combinações de atributos primitivos, através de um processo conhecido como Aprendizado Construtivo.

4.7 Algoritmos de AM mais utilizados na FI

Entre as várias famílias de algoritmos de AM supervisionado existentes para o problema de CT, escolhemos quatro delas para descrever suscintamente nesta Sessão, em razão do seu histórico de aplicação para a atividade de FI, e por terem entre si princípios algoritmicos diferentes. São elas: *Rocchio* (Joachims, 1997; Schapire, Singer e Singhal, 1998; Sebastiani, 2002), *Naive-Bayes* (Domingos e Pazzani, 1996; Joachims, 1997; Mitchell, 1997), *Support Vector Machines* (SVMs) (Vapnik, 1995) e *k-nearest neighbours* (k-NN) (Dasarathy, 1991).

O classificador *Rocchio* é fundamentado no mais conhecido método de aprendizado utilizado em RI: realimentação de relevantes (*relevance feedback*), apresentado por Rocchio (1971). Esse algoritmo foi concebido para atuar na abordagem algébrica para representação de textos, denominada espaço vetorial (Salton e Lesk, 1968), representando os documentos como vetores, de forma que o cálculo de sua similaridade corresponde à semelhança entre os documentos. Cada componente do vetor corresponde a um descritor do documento e seu peso é calculado usando o esquema *tf-idf*. Para o processo de FI, que é reconhecidamente uma atividade de CT com apenas duas categorias mutualmente exclusivas, um vetor protótipo é induzido para a categoria dos exemplo positivos durante a fase de aprendizado. Esse vetor será o próprio classificador que, por meio do cálculo da similaridade entre os vetores, irá filtrar os documentos na fase de teste, descartando os irrelevantes.

Em sua abordagem, Rocchio demonstra, através da Equação (11), como derivar um vetor otimizado de consulta através de suscessivas operações sobre os vetores dos documentos relevantes e não relevantes. Logo em seguida, Robertson e Sparck-Jones (1976) demonstram como ajustar, no modelo probabilístico, o peso individual de um termo baseado na sua distribuição sobre um conjunto de documentos relevantes e não relevantes.

$$QE \;=\; \alpha\, Q \;+\; \frac{\beta}{|R|} \sum_{i=1}^{|R|} R_i \;-\; \frac{\gamma}{|\bar{R}|} \sum_{i=1}^{|\bar{R}|} \bar{R}_i \tag{11}$$

onde:

- QE é o vetor da consulta expandida;

- Q é o vetor da consulta original;
- $|R|$ é o número de documentos relevantes;
- $\left|\bar{R}\right|$ é o número de documentos não relevantes;
- R_i é o vetor do documento relevante i;
- $\bar{R}_i$ é o vetor do documento não relevante i;
- α, β e γ são pesos que determinam a importância dos termos de Q;

Naive-Bayes constitui outro grupo de algoritmos para o aprendizado indutivo com abordagem probabilística, baseada na mesma representação *bag-of-words*. Ele é uma simplificação funcional do classificador ideal Bayesiano, que considera a independência dos atributos dos documentos para gerar um modelo estatístico através da prévia observação do conjunto de testes. Essa abordagem assume que é possível computar a distribuição dos descritores dos documentos previamente associados às categorias. O algoritmo, então, utiliza essa distribuição probabilística para estimar a semelhança que um certo documento tem com uma determinada categoria, formulando assim o seu critério de tomada de decisão.

Naive-Bayes é expressa pela Equação (12), assumindo uma condição de independência entre os descritores. Dessa forma, é computada a probabilidade condicional de um documento D pertencer a uma classe C_i. Essa probabilidade é uma função da frequência com que os descritores de um determinado documento também ocorram em outros documentos cuja categoria já é previamente conhecida.

$$P(C_i|D) \;=\; \frac{P(D|C_i) \;\cdot\; P(C_i)}{P(D)} \tag{12}$$

onde:

- $P(C_i|D)$ é a probabilidade condicional da categoria C_i, dado um documento D;
- $P(D|C_i)$ é a probabilidade condicional do documento D, dada uma categoria C_i;
- $P(C_i)$ é a probabilidade a priori[32] (ou marginal) de C_i;
- $P(D)$[33] é a probabilidade a priori (ou marginal) de D;

[32] no sentido de que não considera nenhuma informação sobre D
[33] atua como uma constante de normalização no teorema de Bayes

Ao ignorar as dependências condicionais entre os termos, é possível determinar $P(D|C_i)$ através do produtório das probabilidades individuais de seus descritores, dada uma categoria C_i. Sendo assim, considerando que um documento D seja formado por um conjunto de descritores t_1 à t_n, a Equação (13) fornece $P(D|C_i)$.

$$P(D|C_i) = \prod_{j=1}^{n} P(t_j|C_i) \tag{13}$$

onde:

- $P(t_j|C_i)$ é a probabilidade condicional do descritor t_j, dada uma categoria C_i;

Recentemente desenvolveu-se uma máquina de aprendizagem chamada Máquina de Vetores-Suporte, ou *Support Vector Machines* (SVMs) (Vapnik, 1995), baseada nos princípios da minimização do risco estrutural, proviniente da Teoria da Aprendizagem Estatística. Essa família de algoritmos vem despertando muito interesse nos últimos anos, sendo atualmente bastante utilizada para a indução de classificadores lineares para um espaço de alta dimensionalidade. Essa máquina faz uso de *kernels* ou mapeamentos, que são funções que mapeiam uma instância qualquer do espaço de entrada em sua correspondente no espaço de alta dimensionalidade. O algoritmo cria um hiperplano ótimo, que maximiza a margem de separação dos dados entre duas classes exclusivas, sejam linearmente separáveis ou não. Ou seja, através dos exemplos de treinamento previamente categorizados, o hiperplano (classificador) de margem maximizada é identificado tal que a distância (margem) entre ele e os exemplos mais próximos é ótima. Os parâmetros do hiperplano de margem maximizada são obtidos por meio de algoritmos complexos, cujo embasamento teórico está centrado na solução de problemas de otimização, conhecidos como *programação quadrática*. Existem vários algoritmos para a resolução dessa classe de problemas, destacando-se a Otimização Mínima Sequencial como a mais conhecida entre eles (Platt, 1999).

Neste trabalho foi utilizada uma implementação popular das SVMs para o experimento aqui conduzido, denominado *SVMLight*[34], que se baseia em uma estratégia de decomposição do problema de otimização em uma série de pequenos problemas, de forma que cada um deles possa ser resolvido mais eficazmente. Abstraímos, entretanto, todo o

[34] svmlight.joachims.org

embasamento teórico matemático-estatístico que fundamenta as SVMs, que compreende os Conceitos Relevantes de Otimização, Produto Interno Kernel e Teoria do Aprendizado Estatístico. A abordagem desses conceitos é demasiadamente complexa e extensa para os nossos propósitos, além de que é possível encontrar na Internet diversos documentos relacionados ao tema, além de livros amplamente recomendados.

Os motivos que justificam a escolha das SVMs como a abordagem adotada nesse trabalho variam desde observações sobre os excelentes resultados obtidos em avaliações realizadas em contextos similares (boa capacidade de generalização), até a constatação de seu alto desempenho em problemas relacionados a espaços de alta dimensionalidade, como é na CT. Sua velocidade de execução representa um ganho significativo no tempo de resposta percebido pelo usuário em sistemas de RI, que constitui uma métrica essencial de avaliação desses sistemas. Também é percebido que as SVMs têm a flexibilidade como uma de suas principais vantagens, podendo adaptar-se a diversos tipos de problemas, entre os quais destacamos a classificação binária, foco deste trabalho.

As SVMs tradicionalmente requerem um novo treinamento completo sempre que há uma alteração no conjunto de treinamento. A reutilização de resultados anteriores, proposta pela técnica da SVMs incrementais, torna os aprendizados sucessivos mais rápidos e também pode reduzir o custo de armazenamento ao descartar os exemplos antigos. O aprendizado incremental é fundamental para ambientes reais de sistemas de FI, onde o usuário muda o foco de um tópico ou aperfeiçoa o seu perfil de busca ao longo do tempo, adicionando ou removendo documentos relevantes.

O classificador *k-nearest neighbours* (k-NN) é baseado na hipótese de que exemplos localizados próximos um dos outros, de acordo com uma métrica de similaridade, provavelmente pertencem a uma mesma classe. Ele também é derivado da regra de *Bayes* e usa o cosseno como métrica de similaridade. $knn(\vec{x})$ denota os índices dos k documentos que possuem os maiores cossenos com o documento para classificar $\vec{x}$.

$$h_{knn}\vec{x} = sign\left(\frac{\sum_{i \in knn(\vec{x})} y_i cos(\vec{x}, \vec{x_i})}{\sum_{i \in knn(\vec{x})} cos(\vec{x}, \vec{x_i})}\right)$$

Exitem outros métodos bastante usados para a atividade de CT, que também são empregados no processo de FI em função do domínio da aplicação e de seus requisitos. São eles:

- *Árvore de Decisão*: O C4.5 é o algoritmo mais popular de árvore de decisão, mostrando-se atraente pela produção de bons resultados em diversos problemas. Ele retorna um nível de confiança ao classificar novos exemplos, que é usado para calcular as métricas de precisão e revocação;

- *Rede Bayesiana*: Um dos problemas do classificador *Naive Bayes* é a hipótese de independência condicional. Usando modelos de rede Bayesianas mais gerais, é possível superar essa limitação. Pesquisas mostram que a construção automática de redes Bayesianas com dependência limitada pode melhorar a performance de previsão.

- *Redes Neurais*: Refere-se ao paradigma de aprendizado conexionista. Este método está relacionado à regressão logística, apesar de que utiliza modelos mais complexos do que os lineares. Como as redes neurais estão muito sujeitas a *overfitting*, é necessário fazer uma pré-seleção de atributos antes de treinar o modelo;

- *Algoritmos de Boosting*: O mais conhecido algoritmo de *boosting* é o AdaBoost (Freund e Schapire, 1996), que combina iterativamente múltiplas hipóteses base (por exemplo, árvores de decisão) usando um modelo linear. *Boosting* também pode ser interpretado como um caso especial de otimização para a maximização de margem do hiperplano, como nas SVMs, com uma função de perda modificada;

- *Aprendizagem de Regras*: Esta abordagem foca em boas estratégias de busca e representações compactas. Um exemplo é o TBL (*Transformation Based Learning*) (Brill, 1995). A vantagem é obter maior interpretabilidade sobre as inferências de classificação, o que caracteriza o aprendizado simbólico.

Nos próximos dois Capítulos apresentaremos os experimentos conduzidos sobre o *CLEF 2006*[35], que consolidam a estratégia adotada para o aumento da eficácia no sistema de RI projetado para o domínio em questão.

[35] www.clef-campaign.org

O *CLEF* é uma iniciativa internacional para avaliação de sistemas de RI para línguas Européias. Tem como objetivo proporcionar o incentivo à colaboração entre comunidades de pesquisadores e o compartilhamento de idéias e resultados. Engloba várias especialidades de sistemas de RI, dentre as quais aquela que restringe o nosso foco de atuação: a atividade de RI *ad-hoc*, monolíngue para o português do Brasil e de Portugal.

A estratégia apresentada consiste de duas etapas: expansão automática de consultas com análise local de sintagmas nominais, visando o aumento da revocação do sistema de RI, imediatamente seguido da aplicação dos FLMs, que visa aumentar a precisão do mesmo sistema. A utilização conjunta dessas duas atividades proporcionou um bom desempenho na avaliação dos resultados do *CLEF 2006* para a atividade em questão, tanto em relação aos resultados obtidos pelo próprio sistema de RI antes do emprego da estratégia conjunta, como em relação aos resultados obtidos por outros sistemas de RI que participaram do mesmo experimento.

5 Expansão de consultas com análise local de sintagmas nominais

Realimentação de relevantes (*relevance feedback*) constitui uma técnica amplamente utilizada para melhorar a performance de sistemas de RI. Seu resultado é diretamente influenciado pela correta escolha dos descritores potencialmente qualificados para exapandir a consulta inicial. Sendo conhecido o poder dos sintagmas nominais (SNs) no papel de descritores com alto poder discriminatório e potencial informativo, neste Capítulo apresentamos uma técnica de análise local para expansão automática de consultas através da utilização de SNs extraídos do conjunto pseudo-relevante. Muito embora a técnica apresentada seja independente da língua, recursos específicos para o português foram utilizados na extração dos SNs através de técnicas de Aprendizado de Máquina.

5.1 O processo de formulação das consultas

A ambiguidade e a imprecisão da linguagem natural são fenômenos frequentes e não chegam a representar um problema para seus falantes. No entanto, quando se trata de processamento computacional, elas são responsáveis pela maioria dos problemas. Os sistemas de RI disponíveis atualmente refletem bem os fenômenos linguísticos dessa natureza, especialmente com relação à expectativa de retorno de respostas coerentes com nossas necessidades de informação. O processo de formulação de consultas constitui um desafio para o sucesso desses sistemas.

Em um sistema de RI, a consulta é definida como o processo de elaboração da necessidade de informação do usuário. Existem diferentes tipos de consulta, sendo todos dependentes do modelo de RI adotado. Por exemplo, as consultas podem ser veiculadas através de protocolos de comunicação em linguagens artificiais ou podem ser elaboradas em linguagem natural, na tentativa de abstrair ao máximo os problemas decorrentes da interação homem-máquina.

Para os modelos de RI mais comumente usados em sistemas Web (Booleano, vetorial e probabilístico), onde o usuário típico é pouco conhecedor do vasto domínio ali indexado, a consulta formulada por palavras-chave destaca-se como a principal linguagem de comu-

nicação homem-máquina (Baeza-Yates e Ribeiro-Neto, 1999). Trata-se de uma linguagem intuitiva, de fácil manipulação e que permite ordenar o conjunto de documentos retornados pela consulta em função de algum critério de relevância. Essa ordenação é uma tarefa difícil de ser conduzida, seja pela inabilidade do usuário em saber articular ou mesmo interpretar eficientemente a sua necessidade de informação, seja pela própria natureza da linguagem humana, caracterizada pela ambiguidade, subjetividade e imprecisão. Entretanto, é nessa comunicação que esses sistemas de RI capturam sua única pista com relação ao suposto objetivo do usuário, havendo quase sempre uma distância semântica entre sua real necessidade e a consulta formulada.

O espaço de descritores compreendido pelo modelo de representação de textos deve corresponder às suas entidades mais significativas e não ambíguas possíveis. Essa intuição é evidenciada por dois fenômenos linguísticos, denominados sinonímia e polissemia. Uma relação de sinonímia existe entre dois descritores morfologicamente diferentes quando ambos possuem mesmo significado semântico. Um descritor é dito polissêmico se, em diferentes contextos, expressa dois ou mais significados distintos. Esses fenômenos são problemáticos para modelos de representação de textos, uma vez que são diretamente responsáveis pelas distorções da revocação e precisão de sistemas de RI, que constituem suas principais medidas de avaliação, de acordo com a expectativa de relevância do usuário. Essa, por sua vez, é caracterizada por uma experiência situacional, subjetiva, cognitiva e dinâmica (Schamber, 1994).

Diante da dificuldade de elaboração de uma consulta bem formulada, o processo de selecionar documentos relevantes a uma necessidade de informação normalmente envolve ciclos interativos entre o usuário e o sistema, compreendendo, na maior parte das vezes, reformulações da consulta inicial.

Uma estratégia para simplificar esse processo consiste em expandir a consulta inicial com termos relacionados, na tentativa de fornecer ao sistema um contexto mais elaborado como consulta, minimizando os problemas inerentes à linguagem humana. Dessa forma, a consulta inicial é apenas uma primeira tentativa de recuperar documentos relevantes. A partir dos documentos por ela retornados, novas consultas são elaboradas com ou sem a intervanção humana, com o intuito de recuperar novos documentos relevantes (revocação), bem como excluir documentos irrelevantes (precisão).

5.2 Expansão de consultas

O processo de reformulação da consulta inicial, denominado expansão de consulta, deve contemplar um critério de seleção para os novos descritores que irão compor a consulta expandida, assim como uma estratégia para recalcular o peso desses descritores juntamente com os da consulta original. Quantos novos descritores selecionar é um problema que deve ser analisado experimentalmente. O importante é que os descritores selecionados reflitam uma carga semântica ou diferencial qualitativa para o contexto onde eles foram identificados, influenciando positivamente o processo.

A expansão da consulta pode ser feita utilizando-se a própria coleção de documentos ou uma base de conhecimento externa à coleção, a exemplo dos tesauros e wordnets[36], que relacionam os termos de um domínio através de medidas de proximidade semântica. Apesar de diversos pesquisadores terem obtido sucesso no emprego de bases externas à expansão de consultas (Gonzalez e Strube de Lima, 2001; Pizzato e Strube de Lima, 2003), o custo de sua obtenção e manutenção geralmente os restringe a aplicações centradas em domínios específicos. Neste trabalho a própria coleção de documentos é a fonte de análise para expansão.

Existem duas abordagens para a expansão de consultas: a) interativa - quando o usuário interage com o sistema fornecendo informações sobre a relevância dos documentos retornados e; b) automática - quando não há interação do usuário no processo de expansão de consulta. Na primeira abordagem, diz-se que há realimentação de relevantes, e normalmente ela é empregada quando a comunidade de usuários é especialista em algum domínio contemplado pela coleção indexada, e quando eles têm disponibilidade para fornecer informações contextuais. A dificuldade do sistema de RI em extrair do usuário evidências contextuais é uma das razões que explica a tendência de se priorizar pesquisas focadas em estratégias automáticas de expansão. Na abordagem automática, diz-se que há pseudo-realimentação de relevantes.

O escopo de documentos analisados para expandir a consulta pode ser global, quando é contemplada toda a coleção de documentos, ou local, quando é analisado apenas um subconjunto da coleção, normalmente os "n" primeiros que já foram retornados em ordem de relevância pela consulta inicial. Foi verificado, em (Baeza-Yates e Ribeiro-Neto, 1999), que

[36] http://wordnet.princeton.edu

na análise global, técnicas de agrupamento de termos são computacionalmente complexas e não produzem resultados sobressalentes, além do que estruturas globais não se adaptam bem ao contexto local de uma consulta, especialmente para coleções genéricas.

A análise local, denominada pseudo-realimentação de relevantes, constitui uma tendência para a automatização do processo de expansão de consultas em coleções de razoável dimensão, independentes de domínio, produzindo comprovadas melhorias na revocação e precisão dos sistemas de RI (Xu e Croft, 2000). Ela tem como vantagem o poder exploratório do contexto local fornecido pela consulta, apresentando-se, dessa forma, mais apropriada do que a análise global. Há muito o que ser estudado sobre técnicas de análise local e, com isso, ela tem se tornado uma promissora tendência de pesquisa. A desvantagem é percebida apenas quando a consulta inicial não retorna ao menos um documento relevante, ou quando no conjunto pseudo-relevante ("top n") exista uma fração suficiente de documentos irrelevantes que introduza descritores ruidosos no processo de expansão, desviando o foco da consulta original (Xu e Croft, 1996; Voorhees e Harman, 1998; Xu e Croft, 2000).

Pseudo-realimentação de relevantes constitui uma solução amplamente utilizada para modificação de consultas desde a década de 70. Rocchio (1971) descreve uma abordagem na qual demonstra como derivar um vetor otimizado de consulta através de suscessivas operações sobre os vetores dos documentos relevantes e não relevantes. Logo em seguida, Robertson e Sparck-Jones (1976) demonstram como ajustar, no modelo probabilístico, o peso individual de um termo baseado na sua distribuição sobre um conjunto de documentos relevantes e não relevantes.

5.3 Identificação de sintagmas nominais

Conforme descrito anteriormente, os SNs são amplamente conhecidos como um conjunto de elementos que faz referências a conceitos, objetos ou fatos do mundo real e, portanto, carregam informações com alto poder discriminatório (Kuramoto, 2002). Essas estruturas linguísticas têm sido amplamente empregadas em diversos problemas computacionais, como a indexação com vocabulário controlado, etiquetagem morfosintática, extração de informações, resolução de co-referência e identificação de relações semânticas. No ex-

perimento conduzido neste Capítulo, eles foram utilizados para o processo de expansão de consultas em sistemas de RI.

O processo de reconhecimento e extração de SNs em textos livres tem evoluído da utilização de programas computacionais que usam motivação simbólica para aqueles que usam motivação estatística. Em parte o motivo se dá pelo amadurecimento das técnicas supervisionadas de AM, à medida que exemplos confiáveis para treinamento dos modelos são disponibilizados.

A experiência considerou apenas os SNs lexicais - aqueles cujo núcleo é um substantivo. Para sua identificação, utilizou-se o sistema de Santos (2005), baseado no algoritmo de aprendizado TBL (*Transformation Based Learning*) (Brill, 1995), que é um dos algoritmos de AM baseados em regras mais bem sucedidos.

Nessa abordagem, o aprendizado é guiado por um corpus de treino que contém exemplos corretamente classificados. O conhecimento linguístico gerado por essa técnica consiste de uma lista ordenada de regras de transformação, que pode ser utilizada para a classificação de novos textos. A lista de regras melhora progressivamente uma classificação inicial abribuída aos itens do corpus de treino.

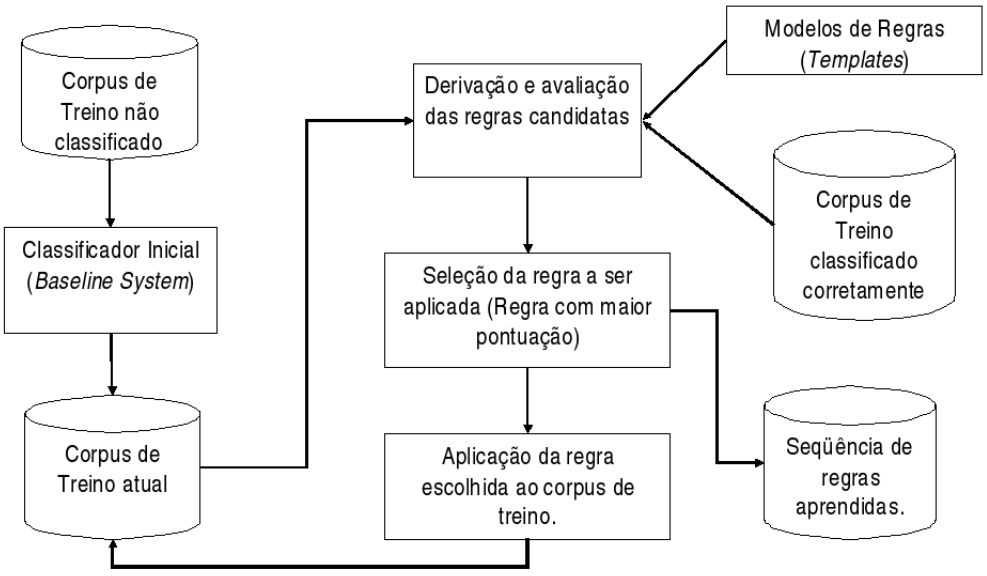


Figura 5: Arquitetura do subsistema de identificação de SNs

Conforme (Santos, 2005), a Figura 5 resume o processo de aprendizado pelo método TBL. Com a utilização de um classificador inicial (*Baseline system*), o início do aprendizado é caracterizado pela atribuição de uma classificação inicial aos itens do corpus de treino. A classificação resultante é comparada com a correta e, para cada erro detectado, todas as regras que o corrigem são geradas a partir da instanciação dos moldes de regras com o contexto do item analisado. A regra que obtiver maior pontuação será selecionada e colocada na lista de regras aprendidas. A regra selecionada é então aplicada ao corpus, e o processo de geração será reiniciado enquanto for possível gerar regras com pontuação acima de um limite especificado. A Figura 6 mostra um exemplo hipotético da entrada e saída do processo de identificação de SNs, onde a saída foi formatada apenas para atender uma necessidade de visualização.

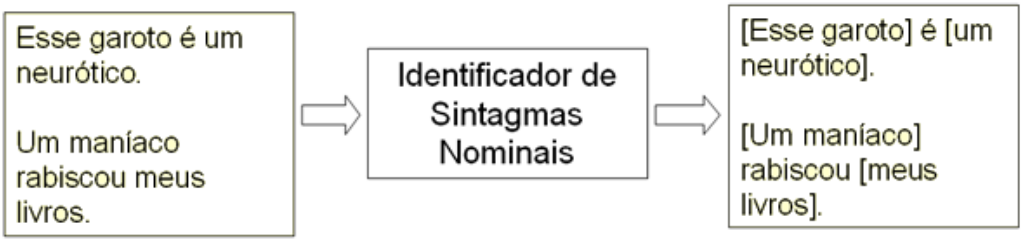


Figura 6: Exemplo de entrada e saída do processo de identificação de SNs (Santos, 2005)

5.4 Descrição do experimento

A proposta de método utilizado neste experimento utiliza a pseudo-realimentação de relevantes para expandir automaticamente a consulta inicial do usuário sem que haja interação deste com o processo de expansão. A análise local foi realizada apenas nos vinte primeiros documentos retornados em ordem de relevância (utilizado-se a equação Okapi BM25 (Robertson e Walker, 1994)) pela consulta inicial. Essa quantidade é empírica e varia em função da própria coleção, do tópico e sua relação com o número de relevantes existente.

5.4.1 Base, tópicos e consultas iniciais

Foi utilizada a base de coleções disponibilizada pelo *CLEF 2006* para a atividade de RI *ad-hoc*, monolíngue, para o português do Brasil e de Portugal. A base é composta por quatro coleções: Folha94, Folha95, Publico94 e Publico95[37], totalizando 210.736 documentos que somam $560Mb$ de texto não estruturado.

Para essa coleção foram disponibilizados 50 tópicos de pesquisa com temas variados, com suas respectivas descrições. Essas descrições foram submetidas a um tratamento manual para formulação do conjunto de consultas iniciais, uma para cada tópico. Cada consulta inicial é composta por uma expressão Booleana sobre os principais descritores identificados na sua respectiva descrição.

Uma operação importante demonimada "proximidade entre termos" foi implementada, uma vez que é útil armazenar também na estrutura de índice a posição absoluta de cada termo do documento. Com isso podemos processar uma operação Booleana, por exemplo, do tipo: $C_n = +d_1 \ -d_2 \ +(d_3 \ \backslash n \ \ d_4)$, onde $n \in Z$. A ausência do operador de proximidade entre d_3 e d_4 é equivalente a "$d_3 \ \backslash 0 \ \ d_4$". Essa consulta C_n busca por todos os documentos que contenham o descritor d_1 e que não contenham o descritor d_2 e que contenham os descritores d_3 e d_4 distantes entre si, no máximo, de n descritores, independentemente da ordem entre eles. Essa operação é fundamental para se trabalhar apenas com os núcleos dos SNs, considerando alguma distância relativa entre eles (independente da ordem), presumindo que essa distância possa evidenciar uma correlação contextual.

A Figura 7 exibe o tópico de número 302, sobre o qual a consulta inicial Q_C302 foi manualmente formulada através de uma linguagem Booleana especialmente desenvolvida para o nosso sistema de RI. Observa-se a utilização de afixos, representado pelo operador "%". Os afixos são úteis para recuperar os termos flexionados no índice, no caso de não utilizar um *stemmer* ou lematizador na respectiva indexação. Essa decisão é útil quando se deseja preservar a função gramatical do termo no índice, para quando for necessário remontar o texto original a partir do mesmo, permitindo um estudo linguístco pós-indexação.

[37] edições completas dos anos de 1994 e 1995 dos jornais PÚBLICO (www.publico.pt) e Folha de São Paulo (www.folha.com.br), compilada pela Linguateca (www.linguateca.pt)

<top>
<num> C302 </num>
<PT-title> Boicotes de consumidores </PT-title>
<PT-desc> Encontrar documentos que descrevam ou discutam o impacto de boicotes por consumidores. </PT-desc>
<PT-narr> Documentos relevantes devem relatar discussões ou pontos de vista sobre a eficácia de boicotes pelos consumidores São também relevantes as questões morais envolvidas nesses boicotes. Apenas boicotes feitos por consumidores são relevantes, boicotes políticos são ignorados.
</PT-narr>
</top>

Q_C302 : all : boicote% consumidor% -"boicote% político%"

Figura 7: Exemplo de formulação da consulta inicial sobre o tópico 302

5.4.2 Pré-processamento da coleção

A coleção necessitou de três etapas distintas de pré-processamento antes de iniciar a indexação. Primeiramente o texto de cada documento foi segmentado em sentenças, estruturando-as uma por linha. Em seguida os termos de cada sentença foram *tokenizados* e analisados morfologicamente. Nessa fase foram feitas as disjunções morfológicas necessárias para o pré-procesamento dos textos, a exemplo de "do = de + o", "àquele = a + aquele", "dentre = de + entre", etc.

Na segunda fase foi feita uma etiquetação sintática (*POS-tagger*) do texto através de Aprendizado de Máquina, utilizando-se o programa *MXPOST* (*Maximum Entropy*, de Ratnaparkhi (1996)), atribuindo para cada palavra uma etiqueta correspondente à sua função gramatical.

O córpus de treino utilizado para induzir o classificador consistiu de 41.883 sentenças extraídas da base Mac-Morpho, do projeto LacioWeb[38].

A terceira fase consistiu da identificação e marcação dos SNs em todas as sentenças de todos os documentos da coleção, enfatizando o núcleo (composto por uma ou mais pala-

[38] http://www.nilc.icmc.usp.br/lacioweb/

vras lexicais) de cada um dos sintagmas. Foi utilizado o algorítmo *TBL* conforme descrito na seção anterior, induzido através de um córpus de treino composto por 4.393 sentenças também extraídas do *Mac-Morpho*. Essas sentenças foram submetidas a uma análise manual (Freitas et al., 2005), para a correta identificação de todos seus SNs. Segundo (Santos, 2005), o processo descrito para identificação dos SNs atinge aproximadamente 87% de *F-measure*.

5.4.3 Indexação com vocabulário controlado

De posse dos textos pré-processados inicia-se a fase de indexação que, por sua vez, também requer operações de pré-processamento comumente utilizadas, a exemplo de *case-folding*, remoção dos acentos e de *stopwords*. Uma vez que os termos estão sintaticamente etiquetados, é possível tomar algumas decisões para a redução da dimensionalidade do espaço de descritores através de indexação com controle de vocabulário. Por exemplo, sequências numéricas não são indexadas a menos que façam parte do núcleo de algum SN. Da mesma forma, tratamos os valores monetários, percentuais, etc. Todos os verbos foram indexados no infinitivo, o que constituiu um dos ganhos mais expressivos em espaço, depois dos valores numéricos. Nomes próprios identificados como entidades nomeadas, por exemplo, foram indexados como descritores formados por múltiplos termos.

5.4.4 Pseudo-realimentação de relevantes

A proposta do método é fazer com que o usuário não precise interagir com o sistema, apesar de que essa interação pode acontecer espontaneamente ao exibir na interface a nova consulta reformulada, antes da sua submissão ao sistema. Nesse ponto o usuário visualiza a nova expressão Booleana, opcionalmente a modifica, e a submete ao sistema procedendo com a iteração expandida. Essa submissão poderia ser completamente automatizada e transparente para o usuário. Contudo, por questões experimentais, achamos válido acompanhar de que forma a consulta foi reformulada e ter o poder de influenciá-la.

Um desafio relevante consiste na exploração de alternativas para identificar quais SNs fornecem os melhores contextos para contribuir eficientemente na reformulação da consulta. Um fator determinante é saber escolher as partes dos documentos das quais serão extraídos

os SNs que atuarão como potenciais candidatos a descritores para a nova consulta. Temos a opção de extraí-los do documento inteiro ou apenas das passagens mais relevantes. Observamos que a coleção trabalhada é uniforme em tamanho e os documentos basicamente se referem apenas a um tópico, com algumas exceções. Optamos por uma terceira alternativa: selecionar apenas aqueles candidatos que estejam próximos da ocorrência sinalizada pela consulta inicial, uma vez que palavras com significados similares tendem a ocorrer em contextos similares (Harris, 1968). O objetivo da escolha é diminuir o ruído causado por descritores que estejam distantes do contexto e, portanto, provavelmente referenciam-se a um tópico diferente. Assim, extraímos todos os SNs da senteça onde foi sinalizada uma ocorrência, bem como os das sentenças imediatamente anterior e posterior. Esses valores são parametrizados no sistema e podem variar entre os experimentos.

Pretende-se selecionar uma quantidade suficiente de novos descritores (também determinada experimentalmente) para compor a nova consulta. Para tanto, temos que atribuir aos SNs pesos que reflitam sua evidência no texto, ou seja, seu destaque sem margem de dúvidas. Para calcular o peso dos SNs, apenas os seus núcleos foram considerados, desprezando-se seus determinantes e modificadores.

O peso de um SN s em um documento d segue a Equação (15), inspirada em Gonzalez (2005).

$$w_{s,d} = f_{s,d} \times \sum_{i=1}^{n} w_{t_i,d} \tag{14}$$

onde:

- $f_{s,d}$ é a frequência de ocorrência de s em d e;
- $w_{t_i,d}$ é o peso do i-ésimo termo t_i do núcleo de s em d;

Cada SN das sentenças escolhidas do documento tem o seu núcleo (normalmente substantivos) segmentado por termos unigrama. Esses termos sofrem um processo de lematização, a fim de proporcionar uma conflação natural entre eles. Os termos lematizados do núcleo dos sintagmas são denominados, apenas no escopo dessa pesquisa, *nucleotídeos*.

O cálculo dos pesos dos nucleotídeos é mais expressivo do que se fosse feito sobre os mesmos termos não lematizados, uma vez que termos morfologicamente distintos e que compartilham um mesmo lema são considerados como uma unidade elementar. Isso os distancia, através do peso, de outros termos semanticamente mais distantes, tornando o processo de seleção mais confiável.

Os nucleotídeos têm seus pesos calculados em função da sua frequência nos SNs do documento. Opcionalmente poderíamos multiplicar esse valor por um fator de ponderação idf, que mede a raridade desse nucleotídeo no conjunto pseudo-relevante (e não em toda a coleção, pois assim descaracteriza a análise local). A frequência de ocorrência do SN s em d é a soma de quantas vezes essa estrutura multi-termos ocorre no documento.

O problema aparece quando duas estruturas com notória proximidade semântica diferem morfologicamente entre si. Por exemplo, "Presidente Fernando Henrique Cardozo" e "Pres. Cardoso, F. Henrrique". Um outro exemplo bem comum na coleção trabalhada (que mistura textos do português brasileiro e de Portugal), é "atividade" e "actividades". A lematização por si só não resolve esses problemas e, para estimar uma medida de similaridade entre esses sintagmas, precisaríamos aplicar funções de *pattern matching* baseadas em *q-grams* e/ou *edit-distance*, por exemplo, o que é perfeitamente exequível em versões futuras do experimento.

Tendo o peso de cada nucleotídeo calculado e a frequência de cada SN no documento d, o peso desse sintagma nesse documento é o produto de sua frequência pelo somatório do peso de seus nucleotídeos. Ordenam-se os SNs do pseudo-conjunto de documentos em ordem decrescente desses pesos e capturam-se os primeiros colocados para compor a consulta expandida.

5.4.5 Avaliação da expansão de consultas

Cada consulta inicial ou expandida elaborada sobre um determinado tópico retorna um conjunto de documentos ordenados por relevância em função daquele tópico. Cada documento retornado é um registro que obedece um formato pré-estabelecido. O conjunto composto por todos os registros agrupados por tópico denomina-se lote (*run*). Cada lote reflete o comportamento do SRI para todos os tópicos disponíveis.

Neste experimento, dois lotes de processamento foram gerados para sua análise comparativa: *i*) NILC01 - sem o uso de expansão de consultas e; *ii*) NILC02 - com uso de expansão de consultas. Os lotes são avaliados pelo programa *trec_eval*[39], que os processa individualmente contra os julgamentos relevantes elaborados por especialistas.

Utilizamos em nosso experimento as métricas tradicionais de avaliação: *i*) *MAP* (*mean average precision*) - que expressa a média da precisão após cada documento relevante ter sido recuperado. Essa métrica enfatiza o quanto antes documentos relevantes são recuperados; *ii*) Precisão - que expressa quantos documentos relevantes foram recuperados em relação ao número de documentos trazidos; *iii*) Revocação - que expressa quantos documentos relevantes foram recuperados em relação ao total.

Apenas 19 dos 50 tópicos (38%) apresentaram ganho de MAP em relação à consulta inicial. Houve empate em apenas 1 tópico que não retornou resultado sem expansão de consulta e, portanto, não haveria como expandi-la. No total, verificou-se que 30 tópicos apresentaram uma perda de MAP em relação à consulta inicial. Isso significa que, apesar da expansão ter retornado mais documentos relevantes na grande maioria dos tópicos, ela também retornou um número muito maior de documentos irrelevantes, pulverizando os relevantes entre eles, prejudicando o *ranking* do conjunto retornado. Isso justifica a perda de precisão em níveis interpolados de revocação.

A métrica MAP para os dois lotes pode ser visualizada, por tópico, no gráfico de barras da Figura (8). O MAP do lote NILC01 é de $35,20\%$, enquanto que para o lote NILC02 é de $29,01\%$. A *precisão* e *revocação* são mapeadas no gráfico de área (9) que analisa o *trade-off* entre a precisão interpolada para cada ponto de revocação padrão, em uma escala percentual, para todos os tópicos.

Foi percebido que quando o SN é uma entidade nomeada (nome próprio, nome de lugar, entidade, etc), a expansão de consulta é bem sucedida. Nesse caso, um peso extra (*boosting*) deveria ser aplicado ao SN para contrabalancear sua decomposição em termos unigramas, que podem se referir a entidades não relacionadas ao sintagma original.

Nenhuma intervenção foi realizada nos parâmetros que regem o comportamento do sistema de RI, enquanto esse processava todos os tópicos do lote NILC02. Após o expe-

[39] Chris Buckley - http://trec.nist.gov/

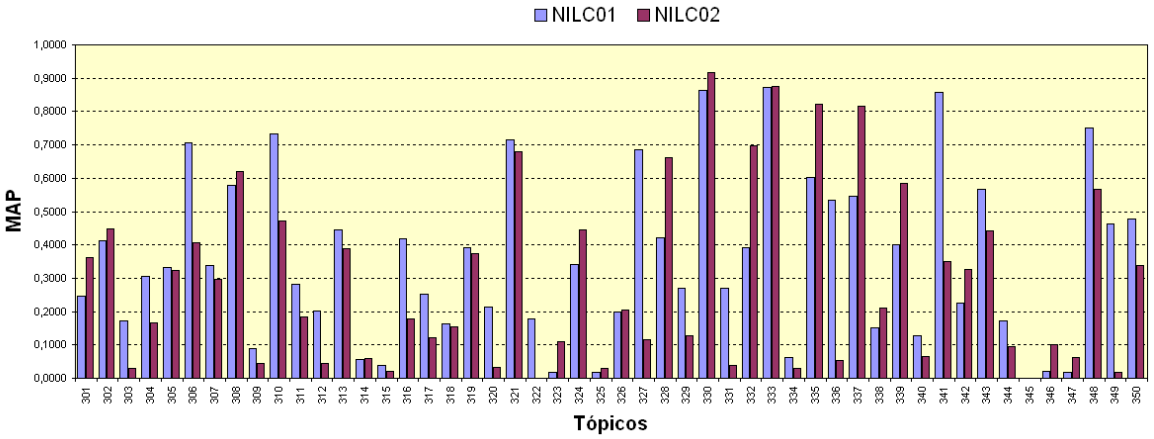


Figura 8: MAP sobre tópicos

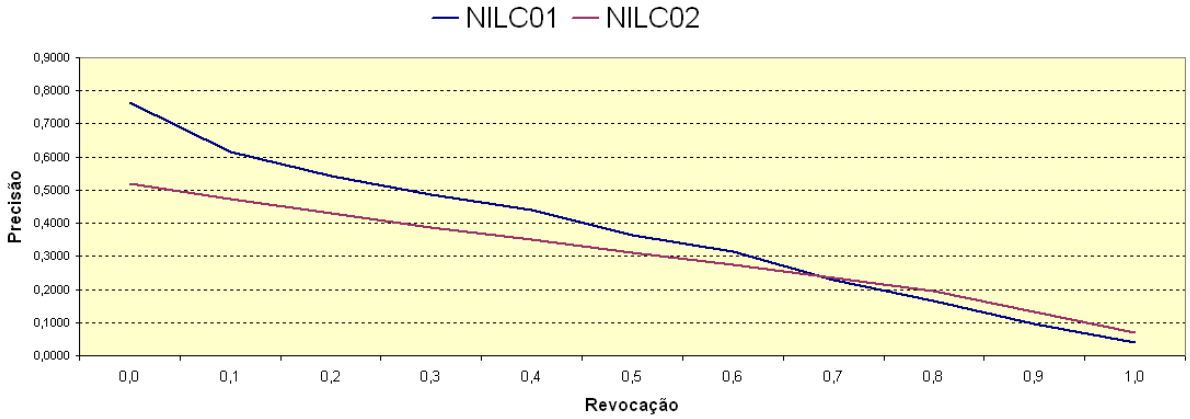


Figura 9: Precisão a cada 10% de revocação obtida

rimento, percebeu-se que a qualidade da consulta inicial é o fator que mais influencia a expansão de consultas. Outros fatores também são responsáveis por influenciar cada tópico, individualmente: *i*) a quantidade de SNs escolhidos; *ii*) a quantidade de sentenças escolhidas para extração dos SNs e; *iii*) a quantidade de documentos do conjunto pseudo-relevante.

Em nossos experimentos, não conseguimos determinar uma relação que explique como estes fatores quantitativos influenciam a expansão de consulta, diferentemente do que aconteceu quando identifica-se a natureza do SN como uma entidade nomeada. Os fatores quantitativos comportam-se como "números mágicos", variando bastante os resultados para cada tópico. Para que se possa analisar esta relação em busca de respostas, é preciso coletar os

resultados de uma grande quantidade de experimentos que correlacionem a natureza dos SNs selecionados com cada fator quantitativo.

Muito embora esse método não tenha apresentado resultados satisfatórios nesse experimento, faz-se necessário experimentar a manipulação individual da consulta expandida para cada tópico, antes de submetê-la ao sistema de RI, a fim de que se possa formular a melhor combinação dos parâmetros do sistema. A observação desse comportamento certamente revelará resultados mais conclusivos a respeito do experiemento.

O alto custo computacional (em complexdade de espaço e tempo) observado em fase de indexação dos documentos permitiu o emprego de recursos linguísticos em estruturas de dados apropriadas para serem usufruídas pelo usuário quando da sua interação com o sistema em fase de busca. O tempo de expansão da consulta acionada em fase de execução, usando conhecimento linguístico previamente indexado, é aceitável (aproximadamente duas vezes maior que o tempo de execução da consulta inicial) e não interfere negativamente na experiência do usuário.

Existem possibilidades de pesquisa em aberto para explorar como outros processos podem ser beneficiados pelo uso de modelos de representação de textos que utilizam conhecimento linguístico envolvendo o uso dos SNs, em especial para a língua portuguesa. No próximo Capítulo é avaliada a influência dessas estruturas em modelos de categorização automática de textos, que são utilizadas para filtrar os documentos irrelevantes, contribuindo para o aumento da eficiência em sistemas de RI.

6 Aplicação dos Filtros Linguisticamente Motivados

Os métodos de acesso à informação sempre estiveram associados, de uma forma ou de outra, a algum esquema de classificação. À medida que a quantidade de material indexado cresce, os sistemas de RI vão se tornando cada vez mais complexos, pois os usuários esperam recuperar todos e apenas aqueles documentos que estejam de acordo com sua espectativa de relevância situacional. A consequência natural desse processo é que a classificação automática de textos como uma atividade de filtragem de informação (FI) desempenhe um papel crítico nos modernos sistemas de RI.

O presente experimento prático objetiva demonstrar como a atividade de FI pode ser aplicada em um fluxo contínuo de documentos retornados por um sistema de RI, aumentando o seu desempenho e, consequentemente, conduzindo o usuário a uma melhor experiência sobre todo o processo de recuperação. Também é objetivo deste experimento analisar o impacto de um modelo de representação de documentos linguisticamente motivado no contexto da categorização automática de textos (CT).

O protótipo foi construído sobre o mesmo ambiente utilizado no experimento do *CLEF 2006* para a atividade *ad-hoc*, monolíngue para o português do Brasil e de Portugal. Conforme detalhado anteriormente, o experimento apresentou um esquema híbrido de indexação utilizando conhecimento estatístico e linguístico, este último fundamentado nos sintagmas nominais. Foi explorado como a atividade de expansão de consulta pode produzir melhor revocação, muitas vezes em detrimento da precisão, sobre as consultas iniciais. Os FLMs atuam nesse novo experimento à medida que os resultados da expansão de consulta são retornados, bloqueando aqueles documentos classificados como falsos-positivos.

6.1 Arquitetura do subsistema de Filtragem de Informação

O subsistema de FI foi acoplado sobre o sistema principal de RI, de maneira que ambos pudessem desfrutar do mesmo modelo de representação de textos, uma vez que a coleção de documentos trabalhada no *CLEF 2006* já tinha sido previamente pré-processada, indexada e estruturada em um sistema gerenciador de banco de dados relacional. Entretanto, essa arquitetura poderia ter sido modularizada de forma independente, uma vez que ambos os

sistemas não necessariamente precisam operar em conjunto, devido à serialização de suas atividades.

O sistema de RI é conhecido como *ad hoc*, pois a coleção de documentos é estática em relação às consultas. Na atividade de FI, o perfil do usuário, que representa a consolidação dos seus interesses de caráter mais duradouro, é estático em relação ao fluxo contínuo de documentos retornados por uma consulta. Dessa forma, o subsistema de FI pode ser projetado para atuar como um módulo local do usuário, por exemplo, através de um *plug-in* do seu navegador, atuando independentemente do sistema de RI que originou o fluxo de documentos.

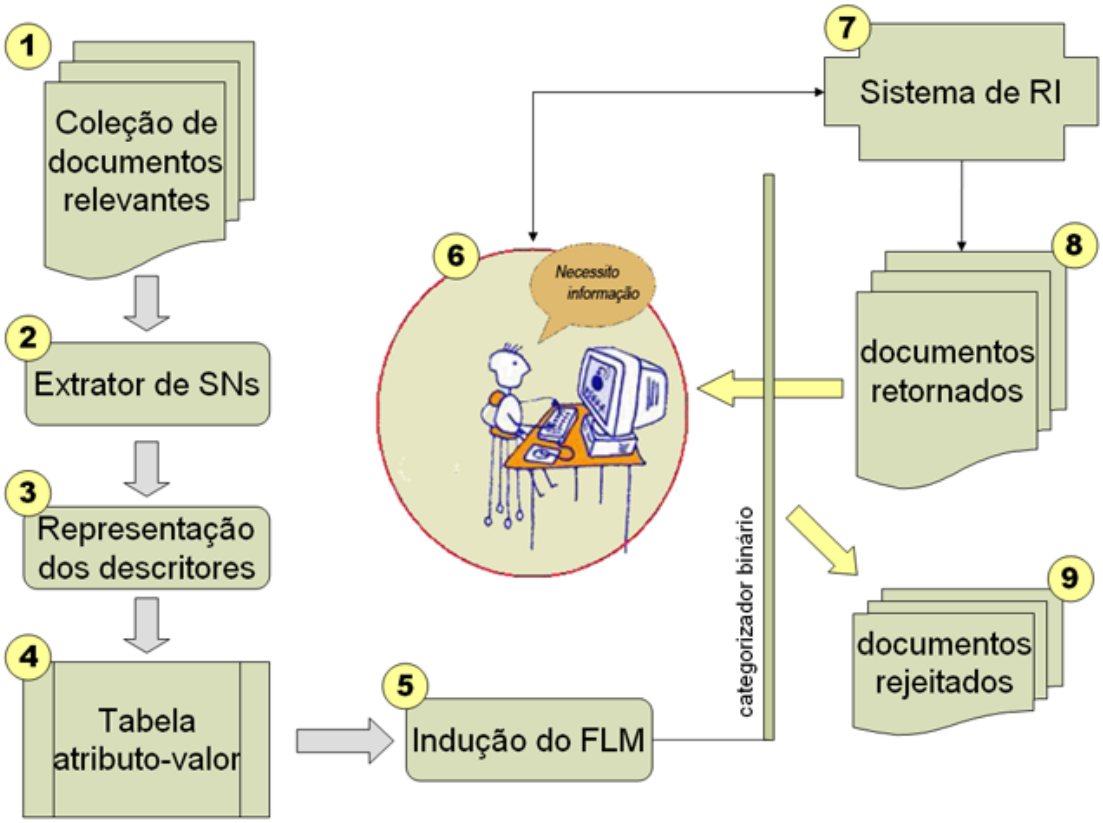


Figura 10: Arquitetura do subsistema de FI em conjunto com o sistema de RI

A arquitetura do subsistema de FI atuando em conjunto com o sistema de RI é apresentado na Figura 10, onde todo o processo encontra-se enumerado em nove etapas distintas, a saber: 1) o usuário mantém em seu domínio um conjunto de documentos relevantes que sintetiza o seu interesse por um determinado tópico de pesquisa; 2) cada

texto é processado pelo identificador de sintagmas nominais citado no Capítulo anterior; 3) o processo de escolha e construção dos descritores, bem como o respectivo cálculo dos pesos que o representam na coleção é detalhado posteriormente nesse Capítulo; 4) construção da tabela atributo-valor, que é a representação estruturada dos documentos que encerram o perfil de busca do usuário; 5) o processo de indução do FLM é conduzido através de técnicas de AM supervisionado, e suas características dependem da família de algoritmos adotada; 6) o usuário inicia o processo iterativo de exploração através da elaboração de sua necessidade de busca; 7) o sistema de RI processa a consulta do usuário e acessa a sua base de dados; 8) o sistema de RI devolve um conjunto de documentos ordenado por algum critério de relevância; 9) o FLM atua sobre os documentos retornados, inferindo decisões sobre a semelhança entre cada documento e o perfil de busca do usuário, bloqueando os documentos irrelevantes.

O processo exploratório do usuário compreende as etapas $[6-9]$, encerrando todas as ações do sistema de RI, incluindo a expansão de consulta e as inferências de similaridade desempenhadas pelos filtros. O processo de indução sobre os julgamentos de relevantes descrito na próxima Seção compreende as etapas $[1-5]$, enquanto que as etapas $[3-4]$ descreve o processo de representação do espaço de descritores, descrito na Seção 6.3.

6.2 Indução sobre os julgamentos de relevantes

As principais métricas de avaliação usadas em sistemas de RI nas últimas três décadas são precisão, revocação e suas curvas de relacionamento. Antes que elas possam ser computadas, é necessário obter os artefatos de um sistema de RI, que compreende uma coleção de documentos, um conjunto de tópicos de consulta e seus respectivos *julgamentos de relevantes*, que são um conjunto de exemplos positivos e negativos de documentos que supostamente melhor representa cada um dos tópicos a serem avaliados.

Conforme exposto no Capítulo 5, o *CLEF 2006* apresentou 50 tópicos de consulta com suas respectivas descrições. Cada tópico representa um interesse específico do usuário sobre um determinado assunto, de onde se pretende capturar sua essência e com ela criar uma estratégia de busca que permita recuperar todos e apenas aqueles documentos relevantes a cada tópico. Após a divulgação dos julgamentos de relevantes para cada tópico do nosso experimento sobre *CLEF 2006*, torna-se viável criar um ambiente que nos permita

induzir um modelo preditivo sobre um subconjunto de exemplos extraídos desses julgamentos. É esperado que o modelo possa generalizar uma decisão binária para cada documento retornado pela atividade de RI, já previamente ordenado em relação à consulta do usuário. Apenas documentos que forem semelhantes ao perfil estático induzido a partir dos julgamentos devem ser mostrados ao usuário, filtrando aqueles que não são relevantes. Assim, os julgamentos foram utilizados nesse experimento para induzir o perfil (filtro) do usuário para um determinado tópico em questão. O subconjunto de documentos que encerra os julgamentos de relevantes para um determinado tópico forma a base de treinamento para a atividade de CT, sob a ótica do AM supervisionado.

Os julgamentos de relevantes são usualmente construídos através de uma técnica denominada *pooling*, onde um conjunto de documentos candidatos (denominado *pool*) é criado através da seleção dos *top-n* (usualmente 100) primeiros documentos classificados como relevantes por todos os sistemas de RI que participam da avaliação sobre um determinado tópico. Existe um problema com a técnica *pooling*, estudado por Zobel (1998), que relata a improbabilidade de que todos os documentos relevantes a um determinado tópico estejam no *pool* e, nesse caso, aqueles que não estiverem poderiam ser classificados erroneamente como não relevantes ao tópico, interferindo nos resultados da avaliação do sistema de RI. Entretanto, essa característica não interfere nos resultados da FI sob a ótica da CT, uma vez que o perfil do usuário (filtro) induzido a partir dos julgamentos de relevantes é contracenado apenas com os documentos retornados pelo sistema de RI, não importando para a avaliação da CT os documentos que não foram recuperados. Em outras palavras, a avaliação do sistema de FI é obtida em relação ao conjunto de documentos retornado pelo sistema de RI (denominado *run*), para aquele tópico em questão.

6.3 Representação do espaço de descritores

É unânime entre os pesquisadores a importância do modelo de representação de dados para toda e qualquer atividade que utilize métodos de acesso à informação. Para sistemas de AM e RI, várias abordagens práticas e teóricas foram bem documentadas na literatura sobre modelos de representação de textos, cada qual na tentativa de obter maior acurácia, envolvendo mais ou menos esforço computacional. Esses modelos variam de unigrama para

aqueles que utilizam alguma estratégia de dependência entre termos, ou ambas, conforme exposto no Capítulo 3.

É suposto que um modelo multitermo adequado possa melhor representar um conceito extraído do texto, muito embora alguns experimentos afirmem que modelos mais sofisticados, em algumas situações, não desempenham melhor que modelos unigrama (Apté et al., 1994). Entretanto, a palavra final sobre o uso de modelos multitermos ainda não foi proferida e existem muitas pesquisas em andamento a seu favor (Sebastiani, 2002).

Seguindo a tendência na adoção de modelos híbridos reportado por Sebastiani (2002), esta pesquisa sobre a atividade de CT procurou por um modelo composto por descritores unigrama em conjunto com unidades de indexação multitermo. Na metodologia apresentada, foi levado em consideração que os conceitos dos textos poderiam ser mais bem representados através dos sintagmas nominais, assim como foi feito para a atividade de expansão de consulta sobre o experimento do *CLEF 2006*. Conforme Kuramoto (2002), essas estruturas constituem um caso especial de relacionamento multitermo porque carregam informações com alto poder discriminatório e potencial informativo.

Nosso modelo de representação com dependência de termos utiliza o conceito de evidência para pesar a relativa confiabilidade de seus descritores. Quanto mais evidência um descritor possuir no texto, maior sua representatividade.

6.4 Indução Construtiva

A seleção de atributos é reconhecidamente um dos problemas mais importantes para o processo de aquisição e descoberta de conhecimento. Indução Construtiva (IC) é o processo de geração de novos atributos potencialmente relevantes para a descrição de um conceito, a partir de combinações de atributos primitivos. No processo de IC esses são denominados operadores.

A produção de novos atributos pode ser conduzida de forma automática ou guiada pelo usuário, podendo demandar muito esforço humano dependendo do domínio do problema. Linguagens de descrição de conceitos inadequadas ou espaços de descritores

estatisticamente mal correlacionados certamente amplificam o problema em si (Lee, 2000), conduzindo à seleção de descritores imprecisos ou excessivamente complexos.

Geralmente o processo de construção de novos descritores é requerido quando os atributos pré-existentes são considerados inadequados, fracamente ou indiretamente relevantes ou foram gerados de forma inapropriada. Contudo, esses atributos primitivos podem ser convenientemente combinados para a geração de novos descritores que melhor representem os conceitos envolvidos. O processo de construção de novos descritores é conhecido como Engenharia de Descritores, Aprendizado Construtivo ou Indução Construtiva (Michalski, 1978).

O objetivo da IC nessa pesquisa é avaliar o impacto de um determinado modelo de representação de textos linguisticamente motivado em algumas atividades de Aprendizado de Máquina (AM) e Recuperação de Informação (RI), em relação aos modelos de representação de textos tradicionais (unigrama). Experimentamos algumas estratégias de combinação de atributos primitivos, selecionados a partir de uma análise fundamentada nos constituintes dos sintagmas nominais, para a geração de novos descritores que melhor representem os conceitos extraídos, e que nos permitam melhor compreender a natureza do domínio explorado.

O método de IC apresentado é automaticamente conduzido pelo sistema e não requer nenhum tipo de intervenção humana. Na metodologia apresentada, o escopo de atributos candidatos para a geração de novos descritores são escolhidos analisando-se os próprios dados do índice do sistema de RI, bem como os dados de treinamento usados pelo sistema de filtragem de informações. Assim, o processo de IC é denominado *indução construtiva orientada a dados*[40] (Bloedorn e Michalski, 1998).

Esses atributos primitivos desempenham o papel de construtores primários no processo de IC, representando as unidades mais atômicas utilizadas para a construção de novos descritores e, dessa forma, definem o espaço amostral para onde são direcionadas as análises apropriadas para a correta seleção e combinação de atributos. Essa análise constitui a base da estratégia adotada para a atividade de IC, e é realizada logo após a etiquetação morfo-sintática do texto e a correpondente identificação dos sintgmas nominais e seus componentes.

[40] data-driven constructive induction

Conforme observado por Lam e Ho (1998), não há esforço que produza melhor acurácia em um processo de RI ou AM, sem o respectivo aumento da complexidade computacional requerida. Na condução de nossos experimentos, não consideramos uma análise que permitisse concluir até onde os esforços linguísticos empreendidos são compensatórios frente aos resultados de desempenho obtidos, muito em função do próprio domínio específico trabalhado nessa aplicação.

Em nosso modelo de espaço de descritores, a relevância de um descritor multitermo é associada a sua representação conceitual, uma vez que o mesmo foi construído a partir de uma estrutura de sintagma nominal que representa um conceito atômico do mundo real. A relevância de um descritor também poderia estar associada a sua probabilidade de distribuição nas classes do conjunto de treinamento, ou mesmo à precisão de classificação obtida. É importante salientar que essas diferentes definições podem classificar o mesmo descritor de diferentes maneiras, em termo de relevância (Lee, 2000).

6.5 Construção de descritores multitermos

Diversas tentativas foram feitas para experimentar diferentes estratégias de construção de descritores multitermos que sejam bem representativos dos conceitos presentes nos texos. Apesar dos esforços empreendidos, certamente não atingimos um modelo insuperável, até mesmo porque o problema da IC é intratável, pois o número de descritores que podem ser derivados é uma função combinatória do número de descritores existentes pelo número de operadores (Lee, 2000).

Entretanto, nós reduzimos as possibilidades de combinação limitando o conjunto de operadores envolvidos. Apenas os núcleos dos sintagmas nominais e seus modificadores são considerados, levando em consideração a ordem natural em que eles aparecem no texto. Uma vez que tais estruturas encontram-se evidenciadas nos textos, algumas heurísticas foram eleitas para combinar o núcleo e modificadores de maneira a produzir melhores resultados no processo de IC, conforme exemplificado abaixo:

1. SN em evidência: {o fascinante [cão Chow] de o admirável [doutor Freud]}

- descritores unigrama: fascinante, cao, chow, admiravel, doutor, freud

- 1^o nível de descritores multitermos: cao_chow, doutor_freud
- 2^o nível de descritores multitermos: cao_chow__doutor_freud
- 3^o nível de descritores multitermos: fascinante__cao_chow, admiravel__doutor_freud

Todos os novos descritores multitermos foram formados concatenando-se os respectivos operadores com o símbolo "_". Foram selecionados três principais níveis de construção de novos descritores: *i*) quando o núcleo de um sintagma nominal for composto por múltiplos termos, esses são concatenados para a formação de um novo descritor; *ii*) se o SN possuir mais de um núcleo, eles devem ser concatenados com "__" para formar um novo descritor; *iii*) os modificadores que orbitarem em torno de seus núcleos devem ser concatenados a esses para a produção de novos descritores. A Figura 11 exemplifica alguns descritores multitermos derivados do processo de IC.

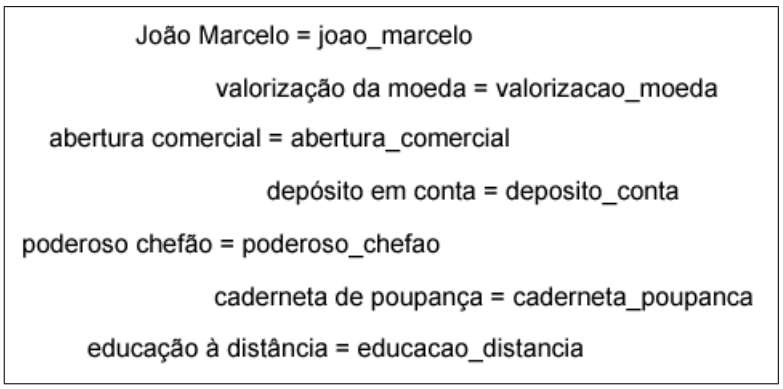


Figura 11: Descritores multitermos derivados do processo de Indução Construtiva

Algumas vezes foi necessário desambiguar a qual núcleo o modificador está associado e resolvemos isso na maioria dos casos utilizando algumas heurísticas simples. Foi percebido que se as atividades de etiquetação e reconhecimento dos sintagmas nominais nos textos fossem conduzidas por um *parser* mais robusto, como o *Palavras* (Bick, 2000), todo o processo traria melhores resultados, com menos ambiguidades para serem resolvidas na hora de escolher as respectivas concatenações. Essa hipótese é baseada na observação de que, apesar das técnicas de Aprendizado de Máquina (AM) serem amplamente reconhecidas para essas atividades, os erros cometidos na identificação dos SNs tendem a se propagar ao longo de todo o processo, até o ponto de observarmos o surgimento de descritores mal formulados, que não representam os conceitos do texto. Entretanto, é esperado que esses falsos-positivos não tenham expressividade estatística no modelo de representação da

coleção e, por isso, sejam descartados em fases posteriores de poda de descritores pelo peso atribuído aos mesmos.

6.6 Esquema de peso dos descritores

Para que seja calculada a confiabilidade de um descritor multitermo em um determinado documento da coleção, foi necessário neste trabalho definir um esquema de peso adequado, que melhor evidencie o seu conceito e importância em relação a todos os outros descritores na coleção. O peso de um descritor multitermo s em um documento d segue a Equação (15).

$$w_{s,d} = f_{s,d} \times \sum_{i=1}^{n} w_{t_i,d} \tag{15}$$

onde:

- $f_{s,d}$ é a frequência de ocorrência de s em d;
- $w_{t_i,d}$ é o peso do i-ésimo termo t_i de s em d; e é definido como:

$$w_{t,d} = \alpha + \beta \times \left(\frac{0.5 \times \log f}{\log f^*} + 0.5 \right) \times \left(\frac{\log \frac{N}{n}}{\log N} \right) \tag{16}$$

que foi introduzido inicialmente pelo sistema de RI probabilístico denominado WIN (Turtle, 1991), onde:

- $w_{t,d}$ é o peso do termo t em d;
- α é uma constante que expressa a probabilidade de relevância (não nula) do termo t na coleção, mesmo que esse termo não ocorra no documento. Seu valor usual é $0,4$, mas nós usamos 0, uma vez que assim obtivemos melhores resultados práticos;
- β é o mesmo que $(1 - \alpha)$ e pesa a contribuição do TF.IDF. Seu valor usual é $0,6$, muito embora usamos 1 em razão dos resultados obtidos;

- f é a frequência do termo t no documento d;
- f^* é a frequência do termo mais frequente no documento;
- N expressa o tamanho da coleção, em número de documentos;
- n é a quantidade de documentos na coleção que contém t

No experimento realizado sobre o *CLEF 2006* foi utilizado um sistema de RI probabilístico e, por questões práticas de se preservar compatibilidade, foi adotado um esquema probabilístico de pesos para os descritores do susbistema de FI. Assim, faz-se necessário satisfazer $0 \leq w_{s,d} \leq 1$, e a Equação (15) não poderia ser utilizada, porque, dados dois pesos w_1 e w_2, sua soma pode exceder 1. Então, é preciso uma função que satisfaça as seguintes propriedades:

(i) $0 \leq f(w_1, w_2) \leq 1$;

(ii) $f(w_1, w_2) \geq \max(w_1, w_2)$;

O peso de uma disjunção "t_1 OU t_2" no modelo probabilístico é dado pela Equação (17);

$$1 - [(1 - w_{t_1,d})(1 - w_{t_2,d})] \tag{17}$$

Para demonstrar a assertiva acima, considere que $P(A \vee B)$ seja a probabilidade de uma operação disjuntiva entre A e B. Então, a seguinte sequência de operações conduz à forma equivalente da Equação (17):

$$\begin{aligned}
&\Rightarrow P(A \vee B) = P(\neg\, \neg(A \vee B));\\
&\Rightarrow P(A \vee B) = 1 - P(\neg(A \vee B));\\
&\Rightarrow P(A \vee B) = 1 - P(\neg A \wedge \neg B));\\
&\Rightarrow P(A \vee B) = 1 - (P(\neg A) \times P(\neg B));\\
&\Rightarrow P(A \vee B) = 1 - ((1 - P(A)) \times (1 - P(B))).
\end{aligned}$$

A Equação (17) satisfaz as propriedades requeridas e, portanto, a mesma foi adotada para substituir a Equação da soma dos pesos. Consequentemente, a Equação (15) pode ser convertida na Equação (18).

$$w_{s,d} = 1 - (1 - S)^{f_{s,d}}, \text{ onde } S = 1 - \prod_{i=1}^{n}(1 - w_{t_i,d}) \tag{18}$$

Isso finalmente conduz à Equação (19).

$$w_{s,d} = 1 - \left(\prod_{i=1}^{n}(1 - w_{t_i,d})\right)^{f_{s,d}} \tag{19}$$

Acreditamos que este cálculo foi uma das contribuições mais expressivas deste trabalho. Foram experimentados alguns modelos para determinar a confiabilidade dos descritores e este foi o que produziu melhores resultados para os algoritmos de AM trabalhados, em especial para as SVMs.

6.7 Tópicos selecionados para o experimento

Para cada um dos 50 tópicos do *CLEF 2006*, um subconjunto dos respectivos julgamentos de relevantes foi utilizado como base de treinamento para induzir os respectivos classificadores binários. Os julgamentos de relevantes são compostos, em média, por um pequeno número de exemplos positivos e negativos, o que inviabilizou as técnicas de AM supervisionado para alguns tópicos, uma vez que se faz necessária uma quantidade razoável de exemplos para caracterizar um bom conjunto de treinamento.

Conforme observado na Figura 12, em média, os julgamentos de relevantes apresentam 403.08 exemplos (variância de 131.49) por tópico, divididos em $53,54$ exemplos positivos (variância de $52,25$) e $349,54$ exemplos negativos (variância de $136,73$). A razão entre os exemplos positivos e negativos é de 21.21% (variância de 24.53%). Dessa forma, foram escolhidos apenas alguns tópicos que apresentassem as maiores quantidades de exem-

plos, muito embora ainda não fossem ideais para o processo de indução supervisionada de categorizadores.

Topic	Total	V	F	%
301	286	28	258	10,85%
302	449	80	369	21,68%
303	439	50	389	12,85%
304	287	58	229	25,33%
305	410	63	347	18,16%
306	237	57	180	31,67%
307	619	33	586	5,63%
308	175	60	115	52,17%
309	414	43	371	11,59%
310	317	131	186	70,43%
311	359	181	178	101,69%
312	544	26	518	5,02%
313	522	215	307	70,03%

Topic	Total	V	F	%
314	387	26	361	7,20%
315	517	82	435	18,85%
316	654	266	388	68,56%
317	423	63	360	17,50%
318	459	11	448	2,46%
319	486	37	449	8,24%
320	534	29	505	5,74%
321	226	40	186	21,51%
322	382	35	347	10,09%
323	664	56	608	9,21%
324	490	110	380	28,95%
325	643	68	575	11,83%
326	526	5	521	0,96%

Topic	Total	V	F	%
327	448	4	444	0,90%
328	275	63	212	29,72%
329	301	12	289	4,15%
330	242	57	185	30,81%
331	249	41	208	19,71%
332	550	6	544	1,10%
333	338	29	309	9,39%
334	258	2	256	0,78%
335	315	36	279	12,90%
336	540	30	510	5,88%
337	255	94	161	58,39%
338	403	11	392	2,81%
339	171	81	90	90,00%

Topic	Total	V	F	%
340	391	13	378	3,44%
341	506	14	492	2,85%
342	321	35	286	12,24%
343	332	70	262	26,72%
344	637	38	599	6,34%
345	398	14	384	3,65%
346	427	25	402	6,22%
347	407	11	396	2,78%
348	463	8	455	1,76%
349	237	40	197	20,30%
350	241	90	151	59,60%
AVG	403,1	53,54	349,5	21,21%
STDV	131,49	52,25	136,73	24,53%

Figura 12: Distribuição dos julgamentos de relevantes sobre os tópicos

Um outro fator considerado para a restrição dos tópicos trabalhados no experimento constitui o problema do desbalanceamento de classes, caracterizado por situações nas quais uma classe é representada por um número muito maior de exemplos em relação à outra. Em alguns domínios, o problema do desbalanceamento de classes pode causar uma performance de classificação subótima (Nitesh et al., 2004). Nesse caso, o classificador superdotado pela classe mais representativa tende a ignorar a classe minoritária. Em nosso domínio, existem muitos tópicos com poucos exemplos positivos em relação aos negativos, e o categorizador tende a aprender a classe negativa melhor que a positiva, produzindo mais erros do tipo falsos-positivos do que falsos-negativos.

Pesquisas recentes têm mostrado que muitos sistemas de aprendizado possuem níveis de insensibilidade com distribuição de classes (Elkan, 2001), a exemplo de *Naive Bayes* e *Árvores de Decisão*. Em nossos experimentos, *Naive Bayes* mostrou-se aparentemente

insensível para o problema, muito embora necessitasse uma avaliação mais precisa, baseado em curvas *ROC* [41].

Existem basicamente três subáreas para tratar o problema do desbalanceamento de classes: *Sampling*, *Feature Selection* e *One Class Learning*. Batista et al. (2004) comparou as várias estratégias, enfatizando um método interessante aplicável quando os conjuntos de dados estão muito desbalanceados ou quando existem muito poucas instâncias da classe minoritária. Todavia, existem algumas evidências enfatizando que métodos para balanceamento de distribuição de classes causam um fraco impacto sobre muitos classificadores. Além disso, métodos de *over-sampling* (balanceamento pela replicação dos exemplos da classe minoritária) podem ampliar o *overfitting*, além de introduzir uma nova atividade computacional ao processo, enquanto que métodos de *under-sampling* (balanceamento pela eliminação de exemplos da classe majoritária) podem eliminar exemplos potencialmente úteis.

O foco deste trabalho é o aprendizado discriminatório (*discrimination-based learning*), ou seja, aquele que utiliza duas classes para a classificação. Uma alternativa para a CT discriminativa é aquela onde apenas os exemplos da classe de interesse são usados para induzir o classificador. Essa abordagem é fundamentada no reconhecimento de uma única classe (*recognition-based learning*), ao contrário da discriminação entre duas classes disjuntas. Nesse caso, a classificação é conduzida através da imposição de um limite (*threshold*) para estimar a similaridade entre a consulta e a classe-alvo (Japkowicz, 2001). Em seu trabalho, Japkowicz observou que, sob certas condições, o aprendizado baseado no reconhecimento de uma classe pode ser superior à abordagem discriminativa. Raskutti e Kowalczyk (2004) demonstraram a optimalidade do método aplicado a classificadores SVMs, em domínios com grandes proporções de classes desbalanceadas. Essa técnica é denominada *extreme re-balancing*, i.e., ignora todos os exemplos da classe minoritária.

A abordagem de aprendizado baseado em uma única classe é recomendada para muitas aplicações do mundo real onde o desbalanceamento de classes é intrínseco ao problema, incluindo a atividade de FI. Nessa aplicação, é esperado que o usuário alimente seu próprio perfil apenas com exemplos positivos que estejam de acordo com seu interesse em algum assunto específico. Entretanto, nosso foco de pesquisa se limitou ao processo de classificação discriminatória e, portanto, preferimos limitar os tópicos do nosso experimento para

[41] Receiver Operating Characterist

aqueles tópicos que possuem uma razão justa entre os exemplos apresentados nos julgamentos de relevantes, que constituem a nossa base de treinamento. Assim, a avaliação da taxa de erro e *F-measure* dos classificadores foi conduzida sobre apenas 8 dos 50 tópicos de pesquisa. Eles são referenciados por seus números: 310, 311, 313, 316, 324, 337, 339 e 350, respectivamente.

6.8 Avaliação dos FLMs sobre o sistema de RI

Um subconjunto dos julgamentos de relevantes para cada um dos 8 tópicos selecionados para a pesquisa foi usado como base de treinamento para induzir 8 classificadores, respectivamente. Para evitar *overfitting* na avaliação do processo de CT, cada documento pertencente à base de treinamento utilizada para induzir o classificador deve ser retirado da base de teste, ou seja, do conjunto de documentos retornado pelo sistema de RI que satisfaz uma consulta sobre um dos tópicos em questão. Por essa razão, a condução do nosso experimento, para cada tópico, só teria validade se fosse utilizada apenas uma porção dos seus julgamentos de relevantes, que já possui um pequeno número de exemplos necessários para proceder com AM supervisionado. Em nossos testes finais, utilizamos 60% dos julgamentos para treinar os classificadores, deixando apenas 40% dos exemplos para servirem de julgamentos contra a base de treinamento.

Foram utilizadas duas famílias diferentes de classificadores para avaliar o nosso modelo híbrido de representação de textos, conforme apresentado, sobre o modelo unigrama tradicional, cujos descritores foram pesados em função da abordagem *TF.IDF*. São elas: *i*) *Naive-Bayes* (NB) e *ii*) Máquinas de vetores-suporte (*Support Vector Machines*) (SVMs). Utilizamos os módulos *AI::Categorizer::Learner::NaiveBayes* e *Algorithm::SVMLight* extraídos do *CPAN* [42], respectivamente.

Foi adotada a técnica de validação cruzada *10-fold stratified cross-validation*[43] para obter a respectiva matriz de confusão e todas as métricas relacionadas: precisão, revocação, taxa de erro e média harmônica entre precisão e revocação, conhecida como *F-measure*.

[42] Comprehensive Perl Archive Network - http://www.cpan.org
[43] preserva a mesma proporção da distribuição de classes entre as partições

As Tabelas 3 e 4 relacionam as médias micro-F1 obtidas pelo processo de validação cruzada, com suas respectivas variâncias. Conforme os resultados, as SVMs obtiveram desempenho aparentemente superior ao classificador *Naive-Bayes* para sete dos oito tópicos selecionados. Naive-Bayes obteve melhor desempenho apenas para o tópico 313. O modelo híbrido de representação de textos também mostrou um desempenho ligeiramente superior ao modelo unitermo, para ambos os classificadores, exceto para o tópico 313 e 337, quando utilizada as SVMs.

Tabela 3: F1-measures obtidas pelo classificador SVM, para ambas representações de texto

	Uniterm	**Desvio Padrão**	**Multiterm**	**Desvio Padrão**
310	81.96%	0.07399	84.93%	0.05095
311	64.04%	0.05995	68.30%	0.08403
313	83.09%	0.07509	82.60%	0.04391
316	75.44%	0.04877	76.94%	0.05144
324	84.40%	0.03864	85.57%	0.04050
337	94.91%	0.06014	92.91%	0.05933
339	67.78%	0.15226	77.19%	0.07910
350	78.75%	0.08998	79.97%	0.07301
Média	**78.80%**	**0.074853**	**81.05%**	**0.06028**

Tabela 4: F1-measures obtidas pelo classificador Naive-Bayes, para ambas representações de texto

	Uniterm	**Desvio Padrão**	**Multiterm**	**Desvio Padrão**
310	77.26%	0.10231	78.39%	0.05284
311	62.47%	0.06945	62.63%	0.06734
313	85.13%	0.05474	86.75%	0.05245
316	70.69%	0.05703	72.03%	0.05255
324	82.40%	0.05128	84.91%	0.04953
337	94.43%	0.04654	95.59%	0.04253
339	66.02%	0.13126	68.73%	0.10996
350	75.42%	0.09335	79.78%	0.06722
Média	**76.73%**	**0.07575**	**78.60%**	**0.05649**

É conhecida a dificuldade de comparação das avaliações obtidas entre diferentes famílias de classificadores. Isso porque eles simplesmente não podem ser diretamente comparados, seja porque as avaliações utilizaram diferentes medidas de desempenho ou mesmo por ter sido utilizado apenas um subconjunto seleto da coleção de documentos. Até mesmo o processo de distribuição aleatória das partições do *cross-validation* pode favorecer uma ou outra circunstância para algum dos classificadores, dentre outras razões (Yang e Liu, 1999).

Assim, alguma análise de significância estatística deve ser conduzida para que se possa julgar qual classificador obteve melhor desempenho sobre as mesmas condições de avaliação. Utilizamos o *t-test* para comparar os classificadores SVM e Naive-Bayes. Um *t-teste* de duas amostras é um teste de hipótese para responder questões sobre a média, onde os dados são coletados de duas amostras aleatóreas, de observações independentes, cada uma com uma distribuição normal subjacente: $N(\mu_i, \sigma_i^2)$, onde $i = 1, 2$. A hipótese nula para as duas amostras *t-test* é: $H_0 : \mu_1 = \mu_2$.

A hipótese nula é testada contra uma das hipóteses alternativas, dependendo da questão proposta:

$$H_1 : \mu_1 \neq \mu_2$$
$$H_1 : \mu_1 < \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Nesse caso, estamos assumindo que as medidas micro-F1 do SVM e Naive-Bayes possuem distribuições normais subjacentes.

Para os testes de hipótese estatística, é preciso escolher um nível de significância, que é uma probabilidade fixada de erroneamente rejeitar a hipótese nula H_0, se ela for de fato verdadeira.

Usualmente, o nível de significância (denotado por α) é escolhido como sendo $0.05 = 5\%$.

O *p-value* (nível de probabilidade) é a probabilidade de erroneamente rejeitar a hipótese nula se ela de fato for verdadeira, calculada a partir das amostras.

O *p-value* é comparado com o nível de significância e, caso seja menor, o resultado é significante. Isto é, se a hipótese nula fosse rejeitada em $\alpha = 0.05$, deveria ser reportado como $p < 0.05$.

Baixos *p-values* sugerem que é pouco provável que a hipótese nula seja verdadeira. Quanto menor for o *p-value*, mais convincente é a rejeição da hipótese nula. Ele indica a força da evidência para, digamos, rejeitar a hipótese nula H_0, em vez de simplesmente concluir "rejeite H_0" ou "não rejeite H_0".

A conclusão final, uma vez que o teste tenha sido efetuado, é sempre dada em termos da hipótese nula. Nós ou "rejeitamos H_0 em favor de H_1" ou "não rejeitamos H_0"; nós nunca concluimos que "rejeitamos H_1", nem que "aceitamos H_1".

Se nós concluirmos "não rejeitamos H_0", isso não necessariamente significa que a hipótese nula é verdadeira, somente sugere que não há evidência contra H_0 em favor de H_1; rejeitar a hipótese nula então sugere que a hipótese alternativa pode ser verdadeira.

Assim, para nosso teste, compararemos as medidas micro-F1 do SVM e Naive-Bayes para coleções unitermo e multitermo. À primeira vista, na média, o micro-F1 do SVM é maior que o do Naive-Bayes em ambas as coleções, então nós faremos um *t-teste* para obter evidência para essa hipótese.

A hipótese nula é:

$$H_0 : \mu_{\text{NB}} = \mu_{\text{SVM}}$$

e a hipótese alternativa é:

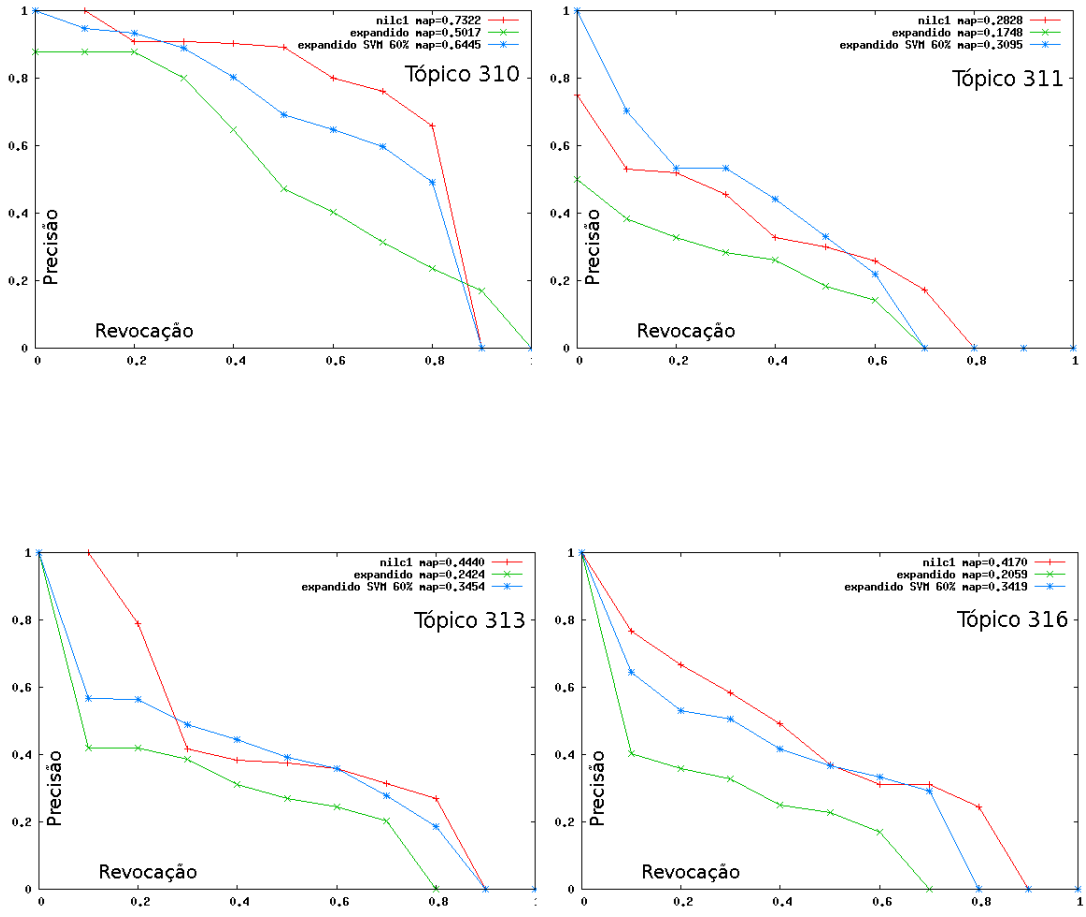
$$H_1 : \mu_{\text{NB}} < \mu_{\text{SVM}}$$

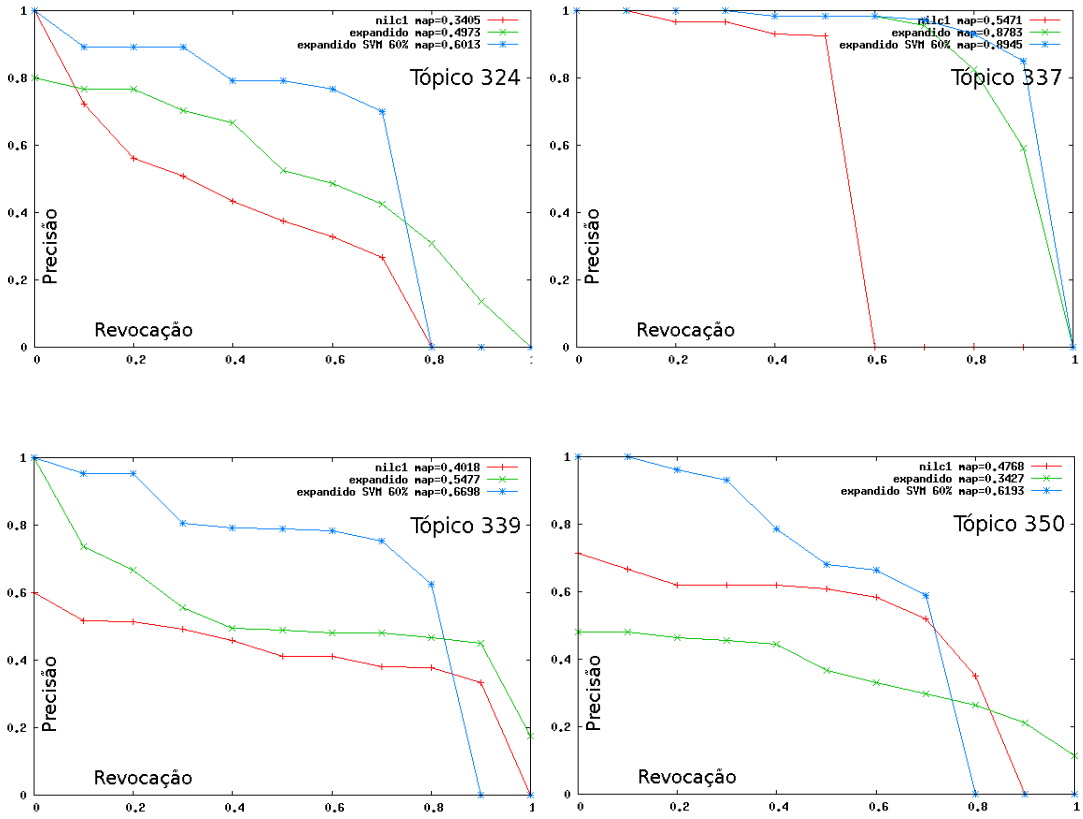
Para uma amostra de 640 micro-F1 para cada modelo, em um cross-validation sobre a coleção unitermo, nós obtivemos um p-value de $0.01858 < 0.05$.

Para outra amostra, de 320 micro-F1 para cada modelo, em um cross-validation sobre a coleção multitermo, obtivemos um p-value de $0.003608 < 0.05$.

Então, dentro de um nível de significância de 0.05, nós podemos rejeitar a hipótese dos micro-F1 do Naive-Bayes e das SVMs serem, na média, iguais, em favor da hipótese de, na média, o micro-F1 do Naive-Bayes ser menor que o das SVMs.

O sistema de RI é usualmente avaliado através de uma métrica conhecida como MAP (*Mean Average Precision*), que é a média da *Precisão Média* (PM) sobre um grupo de consultas, onde PM é a média da precisão após cada documento relevante ter sido recuperado. Nas imagens seguintes, oito plotagens de MAPs são exibidas, uma para cada tópico. Os lotes NILC01 e NILC02 são os mesmos obtidos no *CLEF 2006*, representando as consultas iniciais e expandidas para os tópicos trabalhados, respectivamente. O lote NILC_SVM revela o resultado da filtragem de informação aplicado sobre o lote NILC02, conforme ilustrado na Figura 13. Para cada um dos oito tópicos, a área da curva de MAP obtida com o processo de FI reflete um desempenho superior aos resultados anteriores, i.e., sem o processo de FI.

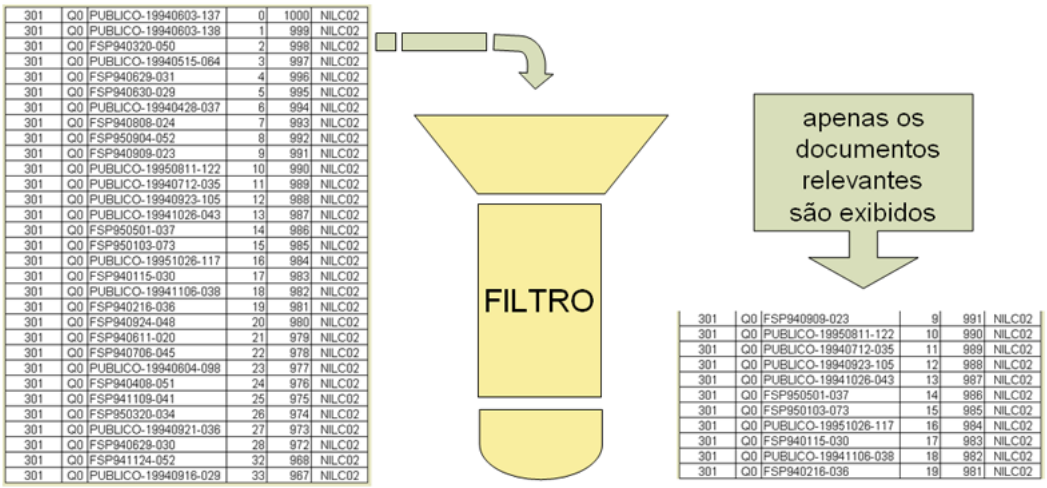




Por conveniência experimental, os sistemas de RI e FI utilizaram o mesmo modelo híbrido de representação de textos. Contudo, o sistema de FI é completamente independente do sistema de RI, conforme já foi anteriormente explicado, de maneira que o subsistema de filtragem poderia ter sido projetado para estruturar os documentos do fluxo à medida que estes são apresentados, já previamente ordenados em função de um critério qualquer do sistema de RI. Este, por sua vez, poderia ter sido estruturado sobre qualquer modelo de representação de textos.

Conforme dito anteriormente, para cada tópico, o classificador escolhido para representar o FLM (através das SVMs) foi induzido utilizando 60% do respectivo julgamento de relevantes, de maneira a evitar que o filtro classifique o mesmo documento que foi usado para treiná-lo, o que causaria um *overfitting* nos resultados. Os julgamentos, por sua vez, apresentam uma distribuição de classes linearmente separáveis razoavelmente balanceada, fazendo com que o filtro aprenda os exemplos positivos tão bem quanto os negativos. Quando o fluxo de documentos representado pelo lote NILC02 modificado (i.e., sem os documentos utilizados para treinar o classificador) é contrastado com o FLM, apenas aqueles

que satisfazem uma decisão de classificação baseado em um critério de similaridade são apresentados ao usuário.



301	Q0	PUBLICO-19940603-137	0	1000	NILC02
301	Q0	PUBLICO-19940603-138	1	999	NILC02
301	Q0	FSP940320-050	2	998	NILC02
301	Q0	PUBLICO-19940515-064	3	997	NILC02
301	Q0	FSP940629-031	4	996	NILC02
301	Q0	FSP940630-029	5	995	NILC02
301	Q0	PUBLICO-19940428-037	6	994	NILC02
301	Q0	FSP940808-024	7	993	NILC02
301	Q0	FSP950904-052	8	992	NILC02
301	Q0	FSP940909-023	9	991	NILC02
301	Q0	PUBLICO-19950811-122	10	990	NILC02
301	Q0	PUBLICO-19940712-035	11	989	NILC02
301	Q0	PUBLICO-19940923-105	12	988	NILC02
301	Q0	PUBLICO-19941026-043	13	987	NILC02
301	Q0	FSP950501-037	14	986	NILC02
301	Q0	FSP950103-073	15	985	NILC02
301	Q0	PUBLICO-19951026-117	16	984	NILC02
301	Q0	FSP940115-030	17	983	NILC02
301	Q0	PUBLICO-19941106-038	18	982	NILC02
301	Q0	FSP940216-036	19	981	NILC02
301	Q0	FSP940924-048	20	980	NILC02
301	Q0	FSP940611-020	21	979	NILC02
301	Q0	FSP940706-045	22	978	NILC02
301	Q0	PUBLICO-19940604-098	23	977	NILC02
301	Q0	FSP940408-051	24	976	NILC02
301	Q0	FSP941109-041	25	975	NILC02
301	Q0	FSP950320-034	26	974	NILC02
301	Q0	PUBLICO-19940921-036	27	973	NILC02
301	Q0	FSP940629-030	28	972	NILC02
301	Q0	FSP941124-052	32	968	NILC02
301	Q0	PUBLICO-19940916-029	33	967	NILC02

301	Q0	FSP940909-023	9	991	NILC02
301	Q0	PUBLICO-19950811-122	10	990	NILC02
301	Q0	PUBLICO-19940712-035	11	989	NILC02
301	Q0	PUBLICO-19940923-105	12	988	NILC02
301	Q0	PUBLICO-19941026-043	13	987	NILC02
301	Q0	FSP950501-037	14	986	NILC02
301	Q0	FSP950103-073	15	985	NILC02
301	Q0	PUBLICO-19951026-117	16	984	NILC02
301	Q0	FSP940115-030	17	983	NILC02
301	Q0	PUBLICO-19941106-038	18	982	NILC02
301	Q0	FSP940216-036	19	981	NILC02

Figura 13: Lote NILC_SVM obtido do processo de FI sobre o lote NILC02

O resultado pós-filtragem é o lote NILC_SVM, que é contrastado com os novos julgamentos de relevantes modificados (sem os documentos utilizados para o treinamento dos filtros). Essa comparação é realizada pelo programa *trec_eval* [44], da mesma maneira que o utilizamos na expansão de consulta realizada no *CLEF 2006*.

[44] Chris Buckley - http://trec.nist.gov/

7 Conclusões e Trabalhos Futuros

Neste trabalho foi proposto um modelo híbrido de representação de textos que utiliza conhecimento linguístico em adição ao conhecimento estatístico em sistemas de RI auxiliado por um processo coadjuvante de pós-filtragem de informação. Para a sua realização, foi empreendido um projeto modular de sistema de CT integrado a um sistema de RI, cujos recursos de PLN empregados fossem suficientes para proporcionar um enriquecimento da experiência do usuário por busca de informações relevantes a diversos tópicos.

O ambiente computacional que abrange o sistema de RI foi avaliado no *CLEF 2006* para a atividade *ad-hoc*, monolíngue, para o português do Brasil e de Portugal. Foi explorado o processo de expansão automática de consultas com análise local de sintagmas nominais. Uma ampla coleção de documentos foi utilizada, juntamente com um conjunto de tópicos de consultas e julgamentos de relevantes.

O processo de Filtragem de Informações (FI) foi realizado através de técnicas de Aprendizado de Máquina (AM). Foi proposto um modelo de representação de documentos linguisticamente motivado pelo uso intensivo dos sintagmas nominais, promovendo a construção de descritores multitermos com alto poder informativo. Esses descritores atuaram em conjunto com o modelo unigrama tradicional para compor a tabelas atributo-valor utilizadas pelos classificadores, proporcionando uma ligeira melhora no desempenho do processo de classificação automática de textos (CT), para ambas as famílias de algoritmos empregadas: SVMs e *Naive-Bayes*.

Um ponto chave no processo de FI foi a seleção de tópicos do *CLEF 2006* que poderiam ser utilizados para o experimento de CT discriminativa, analisando-se quais dos respectivos julgamentos de relevantes disponibilizados poderiam desempenhar o papel de base de treino para os categorizadores. Os critérios de seleção dos tópicos exigiram que *i*) os respectivos julgamentos possuissem uma quantidade mínima de exemplos positivos e negativos, e *i*) que os exemplos de ambas as classes fossem equilibrados em número, para não caracterizar um desbalanceamento. Assim, apenas 8 dos 50 tópicos de pesquisa foram selecionados, muito embora esses ainda não caracterizavam bases de treino ideais em função da quantidade de exemplos para se trabalhar com AM supervisionado. Entretanto, por questões de praticidade, tempo e disponibiliade de recursos, essa era a única forma de

avaliarmos o resultado do processo de FI sobre o sistema de RI, pois esse já havia sido avaliado no *CLEF 2006*. Houve um ganho de $13,13\%$ de MAP, em média, para os oito tópicos selecionados.

Lam e Ho (1998) afirmaram que todo esforço que produza melhor acurácia no processo de CT requer um aumento recíproco da complexidade computacional envolvida no processo. Entretanto, o aumento da complexidade necessária em nosso ambiente ocorreu durante a fase de pré-processamento e indexação, para ambos os sistemas de RI e treinamento dos classificadores. Em modo de busca, a complexidade computacional para a ativiadade de CT é independente da representação de documentos com motivação linguística, uma vez que é preservado o modelo proposicional da CT, visto que os descritores multitermos são apenas novos atributos, comportando-se como outro símbolo qualquer.

Uma vantagem potencialmente interessante na elaboração do espaço de descritores formado por sintagmas nominais seria a possibilidade de empregar técnicas avançadas de reconhecimento de padrões em textos, para que fossem conflacionados certos descritores multitermos, por exemplo, "Presidente Fernando Henrique Cardozo" e "Pres. Cardoso, F. Henrrique", que referem-se ao mesmo descritor. Um outro exemplo bem comum na coleção trabalhada, que mistura textos do português brasileiro e de Portugal, é "atividade" e "actividades". Métricas de similaridade baseadas em *q-grams* e/ou *edit-distance* poderiam ser aplicadas nesses contextos.

Na expansão automática de consulta realizada por realimentação cega de relevantes, apenas 19 dos 50 tópicos (38%) apresentaram ganho de MAP em relação à consulta inicial. No total, verificou-se que 30 tópicos apresentaram uma perda de MAP em relação à consulta inicial. Isso significa que, apesar da expansão ter retornado mais documentos relevantes na grande maioria dos tópicos, ela também retornou um número muito maior de documentos irrelevantes, pulverizando os relevantes entre eles, prejudicando o *ranking* do conjunto retornado. Isso justifica a perda de precisão em níveis interpolados de revocação. Muito embora esse método não tenha apresentado resultados satisfatórios neste experimento, faz-se necessário experimentar a manipulação individual da consulta expandida para cada tópico, antes de submetê-la ao sistema de RI, a fim de que se possa formular a melhor combinação dos parâmetros do sistema. A observação desse comportamento certamente revelará resultados mais conclusivos a respeito do experiemento.

Grande parte das aplicações do mundo real apresenta o problema do desbalanceamento de classes, incluindo a atividade de Filtragem de Informação. Nesses cenários, a abordagem de aprendizado baseado em uma única classe é recomendada, dentre outros métodos citados no Capítulo 6. Muito embora nosso foco de pesquisa limitou-se ao processo de classificação discriminatória e, portanto, preferimos escolher os tópicos que apresentassem uma razão justa entre os exemplos positivos e negativos, trabalhos futuros poderiam explorar estratégias para aqueles tópicos que apresentam o fenômeno das classes desbalanceadas.

Conforme demonstrado, o uso efetivo do PLN em um cenário de CT não está associado a um elevado ganho de acurácia sobre as abordagem tradicionais, mesmo requerendo um aumento substancial da complexidade computacional envolvida. Outrosim, encontra-se relacionado ao aumento de qualidade dos descritores produzidos, proporcionando um melhor entendimento sobre a natureza conceitual das categorias envolvidas. Conhecendo-as melhor, torna-se possível gerenciá-las manualmente para agregar conhecimento externo. Essa situação certamente é útil para ajustar esses sistemas para desempenhar um alto índice de previsibilidade em domínios específicos.

A atividade de CT demonstrou ser um importante instrumento coadjuvante em sistemas de RI, conduzindo o usuário a uma melhor experiência de busca por informações relevantes. No cenário aqui proposto, a CT atuou no processo de pós-filtragem de documentos, bloqueando com satisfatória precisão aqueles cujo conteudo não sejam compatíveis com os interesses mais duradouros do usuário, expressos através do seu perfil de busca.

Enfim, as conclusões apresentadas sugerem a possibilidade de replicar a experiência para muitas aplicações do mundo real, enriquecendo a experiência do usuário na busca incessante por informação relevante.

Referências Bibliográficas

Aires, R. *Uso de marcadores estilísticos para a busca na web em português.* Tese de Doutoramento, Universidade de São Paulo (USP) - Campus de São Carlos, 2005.

Amati, G.; Crestani, F.; Ubaldini, F.; De Nardis, S. Probabilistic learning for information filtering. In: Devroye, L.; Chrisment, C., eds. *Proceedings of RIAO-97, 1st International Conference "Recherche d'Information Assistee par Ordinateur"*, Montreal, CA, 1997, p. 513–530.

Apté, C.; Damerau, F.; Weiss, S. M. Towards language independent automated learning of text categorisation models. In: *Research and Development in Information Retrieval*, 1994, p. 23–30.

Arampatzis, A.; Weide, T.; Koster, C.; Bommel, P. An evaluation of linguistically-motivated indexing schemes. In: *Proceedings of the BCSIRSG '2000*, 2000a.

Arampatzis, A.; Weide, T.; Koster, C.; Bommel, P. Linguistically-motivated information retrieval. In: *Encyclopedia of Library and Information Science*, Marcel Dekker, Inc., New York, Basel, 2000b.

Baclace, P. E. Information intake filtering. In Proceedings of Bellcore Workshop on High-Performance Information Filtering, 1991.

Baeza-Yates, R. Challenges in the interaction of information retrieval and natural language processing. In: *CICLing*, 2004a, p. 445–456.

Baeza-Yates, R. Excavando la web (mining the web, original in spanish). *El profesional de la información (The Information Professional)*, v. 13, n. 1, p. 4–10, 2004b.

Baeza-Yates, R.; Ribeiro-Neto, B. *Modern information retrieval.* Harlow, England: Addison Wesley and ACM Press, 1999.

Batista, G. E. A. P. A.; Prati, R. C.; Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, v. 6, n. 1, p. 20–29, 2004.

Baudisch, P. *Dynamic information filtering.* Tese de Doutoramento, GMD Forschungszentrum Informationstechnik GmbH, Sankt Augustin, iSSN 1435-2699, ISBN 3-88457-399-3., 2001.

Belkin, N. J.; Croft, B. B. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, v. 35, n. 12, p. 29–38, 1992.

Bick, E. *The parsing system palavras: Automatic grammatical analysis of portuguese in a constraint grammar framework.* Tese de Doutoramento, Aarhus University, dr.phil. thesis, 2000.

Bloedorn, E.; Michalski, R. S. Data-driven constructive induction. *IEEE Intelligent Systems*, v. 13, n. 2, p. 30–37, 1998.

Brill, E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, v. 21, n. 4, p. 543–565, 1995.

Carbonell, J. G.; Michalski, R. S.; Mitchell, T. M. An overview of machine learning. In: Michalski, R. S.; Carbonell, J. G.; Mitchell, T. M., eds. *Machine Learning: An Artificial Intelligence Approach*, Berlin, Heidelberg: Springer, p. 3–23, 1984.

Caropreso, M. F.; Matwin, S.; Sebastiani, F. *Statistical phrases in automated text categorization.* Relatório Técnico IEI-B4-07-2000, Pisa, IT, 2001.

Chandrasekar, R.; Srinivas, B. Gleaning information from the web: Using syntax to filter out irrelevant information. 1997.

Church, K. A stochastic parts program and noun phrase parser for unrestricted text. In: *Proceedings of the Second Conference on Applied Natural Language Processing*, 1988, p. 136–143.

Cooper, W. S. Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Trans. Inf. Syst.*, v. 13, n. 1, p. 100–111, 1995.

Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995.

Dasarathy, B. V. *Nearest neighbor pattern classification techniques.* IEEE Computer Society Press, 1991.

Dias, G.; Guilloré, S.; Bassano, J. C.; Pereira-Lopes, J. G. Extraction automatique d'unités lexicales complexes : Un enjeu fondamental pour la recherche documentaire. *Revue T.A.L. (Le traitement automatique des langues) - Traitement automatique des langues pour la recherche d'information*, v. 41, n. 2, 2000.

Domingos, P.; Pazzani, M. J. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In: *ICML*, 1996, p. 105–112.

Elkan, C. The foundations of cost-sensitive learning. In: *IJCAI*, 2001, p. 973–978.

Elkhalifa, L.; Adaikkalavan, R.; Chakravarthy, S. Infofilter: a system for expressive pattern specification and detection over text streams. In: *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, New York, NY, USA: ACM Press, 2005, p. 1084–1088.

Evans, D. A.; Zhai, C. Noun-phrase analysis in unrestricted text for information retrieval. In: *Proceedings of the ACL-96, 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, US, 1996, p. 17–24.

Ferreira, A. B. d. H. *Dicionário aurélio eletrônico - século xxi - versão integral do novo dicionário da língua portuguesa.* Rio de Janeiro - RJ: Editora Nova Fronteira S.A., 1999.

Freitas, M. C.; Garrãoo, M.; C., O.; Santos, C. N.; Silveira, M. A anotação de um corpus para o aprendizado supervisionado de um modelo de sn. In: *Proceedings of the III TIL / XXV Congresso da SBC*, São Leopoldo - RS, 2005.

Freund, Y.; Schapire, R. E. Experiments with a new boosting algorithm. In: *International Conference on Machine Learning*, 1996, p. 148–156.

Goldberg, D.; Nichols, D.; Oki, B. M.; Terry, D. Using collaborative filtering to weave an information tapestry. *cacm*, v. 35, n. 12, p. 61–70, 1992.

Gonzalez, M. *Termos e relacionamentos em evidência na recuperação de informação.* Tese de Doutoramento, Universidade Federal do Rio Grande do Sul (UFRGS), 2005.

Gonzalez, M. A. I.; Strube de Lima, V. L. Recuperação de informação e expansão automática de consulta com thesaurus. In: *XXVII Conferência Latinoamericana de Informática (CLEI'2001)*, Mérida, Venezuela, 2001, p. 1–10.

Harris, Z. *Mathematical structures of language*. New York - USA, 1968.

Houaiss, I. A. *Dicionário eletrônico houaiss da língua portuguesa*. Rio de Janeiro - RJ: Editora Objetiva Ltda., 2002.

Iyer, R. D.; Lewis, D. D.; Schapire, R. E.; Singer, Y.; Singhal, A. Boosting for document routing. In: Agah, A.; Callan, J.; Rundensteiner, E., eds. *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management*, McLean, US: ACM Press, New York, US, 2000, p. 70–77.

Jackson, P.; Moulinier, I. *Natural language processing for online applications: Text retrieval, extraction and categorization*. John Benjamins Publishing Co, 2002.

Japkowicz, N. Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning*, v. 42, n. 1/2, p. 97–122, 2001.

Joachims, T. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: Fisher, D. H., ed. *Proceedings of ICML-97, 14th International Conference on Machine Learning*, Nashville, US: Morgan Kaufmann Publishers, San Francisco, US, 1997, p. 143–151.

Joachims, T. *Learning to classify text using support vector machines*. Dordrecht, NL: Kluwer Academic Publishers, 2002.

Katz, S. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, v. 2, n. 1, p. 15–60, 1996.

Kjersti, A. A survey on personalized information filtering systems for the world wide web. 1997.

Kudo, T.; Matsumoto, Y. Chunking with support vector machines. Pittsburgh, PA - USA: In NAACL-2001. Language technologies 2001 - Proceedingsof the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, 2001, p. 192–199.

KURAMOTO, H. Sintagmas nominais: uma nova proposta para a recuperação de informação. *DataGramaZero - Revista de Ciência da Informação*, v. 3, n. 1, 2002.

LAM, W.; HO, C. Y. Using a generalized instance set for automatic text categorization. In: *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA: ACM Press, 1998, p. 81–89.

LAM, W.; MUKHOPADHYAY, S.; MOSTAFA, J.; PALAKAL, M. Detection of interest shifts for personalized information filtering. 1996, p. 317–324.

LEE, C.; LEE, G. G. Probabilistic information retrieval model for a dependency structured indexing system. *Inf. Process. Manage.*, v. 41, n. 2, p. 161–175, 2005.

LEE, H. D. *Seleção e construção de features relevantes para o aprendizado de máquina.* Dissertação de Mestrado, São Carlos - SP, 2000.

LEWIS, D. D. An evaluation of phrasal and clustered representations on a text categorization task. In: *SIGIR*, 1992, p. 37–50.

LIU, S.; LIU, F.; YU, C. T.; MENG, W. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: *SIGIR*, 2004, p. 266–272.

LOBATO, L. M. P. *Sintaxe gerativa do português: da teoria padrão à teoria de regência e ligação.* Belo Horizonte - MG: Editora Vigília, 1986.

LOSEE, R. M. Minimizing information overload: the ranking of electronic messages. *J. Inf. Sci.*, v. 15, n. 3, p. 179–189, 1989.

LUHN, H. P. A business intelligence system. *j-IBM-JRD*, v. 2, p. 314–319, 1958.

MACSKASSY, S. A. New techniques in intelligent information filtering. 2003.

MICHALSKI, R. S. Pattern recognition as knowledge-guided computer induction. Tech. Report 927, 1978.

MILLER, D. R.; LEEK, T.; SCHWARTZ, R. M. A hidden markov model information retrieval system. In: *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, Berkeley, US, 1999, p. 214–221.

Miller, G. A.; Fellbaum, C. Introduction to WordNet: An On-line Lexical Database*. *Int J Lexicography*, v. 3, n. 4, p. 235–244, 1990.

Miorelli, S. T. *Ed-cer: Extração do sintagma nominal em sentenças em português.* Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre - RS, 2001.

Mitchell, T. M. *Machine learning.* New York: McGraw-Hill, 1997.

Nanas, N.; Uren, V.; Roeck, A. A comparative evaluation of term weighting methods for information filtering. 2003.

Nanas, N.; Uren, V.; Roeck, A. Nootropia: a user profiling model based on a self-organising term network. 2004.

Navarro, G.; Baeza-Yates, R.; Arcoverde, J. M. A. Matchsimile: a flexible approximate matching tool for searching proper names. *J. Am. Soc. Inf. Sci. Technol.*, v. 54, n. 1, p. 3–15, 2003.

Navarro, G.; Raffinot, M. *Flexible pattern matching in strings - practical on-line search algorithms for texts and biological sequences.* Cambridge University Press, iSBN 0-521-81307-7. 280 pages., 2002.

Ngai, G.; Yarowsky, D. Rule writing or annotation: Cost-efficient resource usage for base noun phrase chunking. 2000.

Ngu, D. S.; Wu, X. Sitehelper: A localized agent that helps incremental exploration of the world wide web. In: *www97*, 1997.

Nitesh, V. C.; Japkowicz, N.; Kotcz, A. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations*, v. 6, n. 1, p. 1–6, 2004.

Oard, D. W.; Marchionini, G. *A conceptual framework for text filtering process.* Relatório Técnico CS-TR-3643, 1996.

Perini, M. A. *Para uma nova gramática do português.* São Paulo - SP: Editora Ática, 2000.

Pickens, J.; Croft, B. An exploratory analysis of phrases in text retrieval. 2000.

Pizzato, L. A. S.; Strube de Lima, V. L. Evaluation of a thesaurus-based query expansion technique. In: Mamede, N. J.; Baptista, J.; Trancoso, I.; Nunes, M. V., eds. *Proceedings of the 6th Workshop on Computacional Processing of the Portuguese Language - Written and Spoken. Lecture Notes in Computer Science 2721*, Universidade do Algarve-FCHS, Faro, Portugal.: Springer-Verlag, 2003, p. 251–258.

Platt, J. C. Fast training of support vector machines using sequential minimal optimization, p. 185–208. 1999.

Ramshaw, L.; Marcus, M. Text chunking using transformation-based learning. In: Yarovsky, D.; Church, K., eds. *Proceedings of the Third Workshop on Very Large Corpora*, Somerset, New Jersey: Association for Computational Linguistics, 1995, p. 82–94.

Raskutti, B.; Kowalczyk, A. Extreme rebalancing for svms: a case study. *SIGKDD Explorations*, v. 6, n. 1, p. 60–69, 2004.

Ratnaparkhi, A. A maximum entropy part-of-speech tagger. University of Pennsylvania, USA, 1996.

Rich, E. *Artificial intelligence*. New York: McGraw-Hill, 1983.

Robertson, S. E. The probability ranking principle in IR. *Journal of Documentation*, v. 33, p. 294–304, 1977.

Robertson, S. E.; Sparck-Jones, K. Relevance weighting of search terms. *J. Amer. Soc. for Information Sci.*, v. 27, p. 129–146, 1976.

Robertson, S. E.; Walker, S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: *SIGIR*, 1994, p. 232–241.

Rocchio, J. Relevance feedback in information retrieval. In: Salton, G., ed. *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice Hall, p. 313–323, 1971.

Sahami, M. *Using machine learning to improve information access*. Tese de Doutoramento, Computer Science Department, Stanford University, 1998.

Salton, G.; Buckley, C. *Term weighting approaches in automatic text retrieval*. Relatório Técnico, Ithaca, NY, USA, 1987a.

Salton, G.; Buckley, C. *Term weighting approaches in automatic text retrieval.* Relatório Técnico, Ithaca, NY, USA, 1987b.

Salton, G.; Buckley, C.; Yu, C. T. An evaluation of term dependence models in information retrieval. In: *SIGIR*, 1982, p. 151–173.

Salton, G.; Lesk, M. E. Computer evaluation of indexing and text processing. *J. ACM*, v. 15, n. 1, p. 8–36, 1968.

Salton, G.; McGill, M. J. *Introduction to modern information retrieval.* New York, NY, USA: McGraw-Hill, Inc., 1986.

Santos, C. N. *Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro.* Dissertação de Mestrado, Instituto Militar de Engenharia IME, Rio de Janeiro - RJ, 2005.

Schamber, L. Relevance and information behavior. *Annual Review of Information Science and Technology (ARIST)*, v. 29, p. 3–48, 1994.

Schapire, R. E.; Singer, Y.; Singhal, A. Boosting and rocchio applied to text filtering. In: Croft, W. B.; Moffat, A.; van Rijsbergen, C. J.; Wilkinson, R.; Zobel, J., eds. *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, Melbourne, AU: ACM Press, New York, US, 1998, p. 215–223.

Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys*, v. 34, n. 1, p. 1–47, 2002.

Sheth, B. D. *A learning approach to personalized information filtering.* Dissertação de Mestrado, 1994.

Simon, H. A. Search and reasoning in problem solving. *Artificial Intelligence*, v. 21, n. 1-2, p. 7–29, 1983.

Smeaton, A. Using nlp or nlp resources for information retrieval tasks. 1997.

Smeaton, A. F. Natural language processing and information retrieval. *Information Processing and Management*, v. 26, n. 1, p. 19–20, 1990.

SMEATON, A. F. Using NLP or NLP resources for information retrieval tasks. In: STRZALKOWSKI, T., ed. *Natural Language Information Retrieval*, v. 7 de *Text, Speech and Language Technology*, Dordrecht/Boston/London: Kluwer Academic Publishers, p. 99–111, 1999.

SOUZA, R. R.; ALVARENGA, L. Um projeto de metodologia para escolha automática de descritores para textos digitalizados utilizando sintagmas nominais. In: *XXIV Congresso da Sociedade Brasileira de Computação*, Salvador - BA, iI Til - Workshop de Tecnologia da Informação e da Linguagem Humana, 2004.

SPARCK-JONES, K.; WILLETT, P. *Readings in information retrieval*. San Francisco, California, USA: Morgan Kaufmann Publishers, Inc., 1997.

STRZALKOWSKI, T. Natural language information retrieval. *Inf. Process. Manage.*, v. 31, n. 3, p. 397–417, 1995.

TAURITZ, D. R.; KOK, J. N.; SPRINKHUIZEN-KUYPER, I. G. Adaptive information filtering using evolutionary computation. *Information Sciences*, v. 122, n. 2-4, p. 121–140, 2000.

TSUI, E.; GARNER, B. J.; STAAB, S. The role of artificial intelligence in knowledge management. *Knowledge Based Systems*, v. 13, n. 5, p. 235–239, 2000.

TURTLE, H. R. *Inference networks for document retrieval*. Tese de Doutoramento, 1991.

VAPNIK, V. N. *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

VILARES, J.; BARCALA, F. M.; ALONSO, M. A. Using syntactic dependency-pairs conflation to improve retrieval performance in spanish. In: *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, London, UK: Springer-Verlag, 2002, p. 381–390.

VOORHEES, E. M. Using wordnet to disambiguate word senses for text retrieval. In: *SIGIR*, 1993, p. 171–180.

VOORHEES, E. M.; HARMAN, D. The text retrieval conferences (trecs). In: *Proceedings of a workshop on held at Baltimore, Maryland*, Morristown, NJ, USA: Association for Computational Linguistics, 1998, p. 241–273.

VOUTILAINEN, A. Nptool, a detector of english noun phrases. 1993.

WERMTER, J.; FLUCK, J.; STROETGEN, J.; GEISSLER, S.; HAHN, U. Recognizing noun phrases in biomedical text: An evaluation of lab prototypes and commercial chunker. In: *First International Symposium on Semantic Mining in Biomedicine (SMBM) - 10th -13th April*, 2005.

WILCOX, L. C. *Knowledge management and its integrative elements*. Boca Raton, FL, USA: CRC Press, Inc., 1997.

XU, J.; CROFT, W. B. Query expansion using local and global document analysis. In: *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA: ACM Press, 1996, p. 4–11.

XU, J.; CROFT, W. B. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, v. 18, n. 1, p. 79–112, 2000.

YANG, Y.; LIU, X. A re-examination of text categorization methods. In: *22nd Annual International SIGIR*, Berkley, 1999, p. 42–49.

ZHAI, C. Fast statistical parsing of noun phrases for document indexing. 1997.

ZIPF, G. *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge, MA, 573 pp., 1949.

ZOBEL, J. How reliable are the results of large-scale information retrieval experiments? In: *Research and Development in Information Retrieval*, 1998, p. 307–314.