UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ANDRÉ PINTO GERALDO

Aplicando Algoritmos de Mineração de Regras de Associação para Recuperação de Informações Multilíngues

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dra. Viviane Pereira Moreira
Orientadora

Carlos Alberto Heuser
Co-orientador

Porto Alegre, dezembro de 2009.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Geraldo, André P.

Aplicando Algoritmos de Mineração de Regras de Associação para Recuperação de Informações Multilíngues / André Pinto Geraldo – Porto Alegre: Programa de Pós-Graduação em Computação, 2009.

78 f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2009. Orientador: Viviane Pereira Moreira;

1.Recuperação de informação multilíngue 2.Recuperação de informação 3.Mineração de dados. 4.Regras de associação

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Vice-Reitor: Prof. Rui Vicente Oppermann
Pró-Reitor de Pós-Graduação: Prof. Aldo Bolten Lucion
Diretor do Instituto de Informática: Prof. Flávio Rech Wagner
Coordenador do PPGC: Prof. Álvaro Freitas Moreira
Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

CLEF	Cross Language Evaluation Forum
MD	Mineração de dados
RAs	Regras de Associação
RI	Recuperação de informações
RI-ML	Recuperação de Informações Multilíngues
SGML	Standard Generalized Markup Language
SRI	Sistema de recuperação de informações
TREC	Text Retrieval Conference

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

RESUMO

Este trabalho propõe a utilização de algoritmos de mineração de regras de associação para a Recuperação de Informações Multilíngues. Esses algoritmos têm sido amplamente utilizados para analisar transações de registro de vendas. A ideia é mapear o problema de encontrar associações entre itens vendidos para o problema de encontrar termos equivalentes entre idiomas diferentes em um corpus paralelo. A proposta foi validada por meio de experimentos com diferentes idiomas, conjuntos de consultas e corpora. Os resultados mostram que a eficácia da abordagem proposta é comparável ao estado da arte, ao resultado monolíngue e à tradução automática de consultas, embora este utilize técnicas mais complexas de processamento de linguagem natural. Foi criado um protótipo que faz consultas à Web utilizando o método proposto. O sistema recebe palavras-chave em português, as traduz para o inglês e submete a consulta a diversos *sites* de busca.

Palavras-Chave: Recuperação de informações, recuperação de informações multilíngues, regras de associação.

Cross-Language Information Retrieval using Algorithms for mining Association Rules

ABSTRACT

This work proposes the use of algorithms for mining association rules as an approach for Cross-Language Information Retrieval. These algorithms have been widely used to analyze market basket data. The idea is to map the problem of finding associations between sales items to the problem of finding term translations over a parallel corpus. The proposal was validated by means of experiments using different languages, queries and corpora. The results show that the performance of our proposed approach is comparable to the performance of the monolingual baseline and to query translation via machine translation, even though these systems employ more complex Natural Language Processing techniques. A prototype for cross-language web querying was implemented to test the proposed method. The system accepts keywords in Portuguese, translates them into English and submits the query to several web-sites that provide search functionalities.

Keywords: Information retrieval, cross-language information retrieval, association rules.

1 INTRODUÇÃO

A Recuperação de Informações (RI) é a ciência que estuda a forma de representação, armazenamento, organização e acesso à informação, objetivando um fácil acesso aos dados de interesse do usuário (BAEZA-YATES; RIBEIRO-NETO, 1999). Os sistemas de RI datam das primeiras aplicações catalográficas surgidas na era digital. Esses primeiros sistemas possuíam como objetivo cadastrar entes do mundo real tais como livros em bibliotecas. Desses objetos eram armazenados apenas metadados ou informações descritivas. Porém, com o avançar da tecnologia e barateamento do armazenamento, tornou-se possível ter coleções de grande volume de dados armazenados em computadores, e não apenas seus metadados. Ao mesmo tempo, as necessidades de obtenção de informação evoluíram e com isso surgiram novos requisitos que não poderiam mais ser atendidos pelos modelos tradicionais de recuperação de dados. Para resolver esse problema, propôs-se a RI.

Atualmente a RI está na vida diária das pessoas por meio dos sistemas de busca, sejam estes de buscas na Internet, em lojas, bibliotecas, mapas digitais, coletâneas de mídias etc. Hoje calcula-se que 95% da informação produzida no mundo não esteja estruturada em bancos de dados (FAVARO; VIEIRA, 2008). Isso torna essas informações inacessíveis ou, em alguns casos, pouco eficientes às propostas tradicionais de recuperação de dados.

A Recuperação de Informações Multilíngues (RI-ML) é uma subárea da RI na qual documentos em um idioma são buscados em resposta a uma consulta formulada em outro idioma, como, por exemplo, a recuperação de documentos em inglês em resposta a uma consulta em português. A principal motivação para a RI-ML é a crescente necessidade de explorar documentos em línguas estrangeiras. Essa necessidade vem crescendo dramaticamente com a disseminação da Internet, que removeu a distância física entre o usuário e a informação. Contudo, a barreira da linguagem ainda precisa ser removida.

Uma das principais aplicações para RI-ML são as consultas à *Web*. Sendo a Internet um repositório multilíngue, o acesso à informação disponível em outros idiomas depende da fluência do usuário nesses idiomas. Por meio da RI-ML pode-se diminuir essa necessidade, permitindo que sejam feitas consultas em um idioma e elas retornem resultados em outras línguas. Como outra aplicação, pode-se citar as bibliotecas digitais nas quais documentos podem ser armazenados em diversos idiomas. Além disso, as bibliotecas digitais são excelentes sistemas para testes, uma vez que são ambientes mais controlados do que a *Web*. Além dessas, podemos citar como aplicações Mercados Comuns (como a União Europeia e o Mercosul), países multilíngues (como a Suíça e o Canadá) e empresas Multinacionais, que possuem grande necessidade de intercâmbio de informações multilíngues.

A barreira da linguagem pode ser exemplificada com a língua portuguesa. O idioma português é o sexto mais falado no mundo, possuindo aproximadamente 230 milhões de falantes distribuídos por nove países onde é língua oficial (CPLP, 2007). A ampla maioria dessa população possui capacidade de formular apenas consultas em português, ficando

restrita a buscas em apenas 1,4% da Internet que está disponível em português[1]. Isso é bastante limitado se comparado aos 68,4% disponíveis em inglês. A isso acrescenta-se que conteúdos não dependentes de idiomas, como imagens, figuras e fotos, possuem seus metadados comumente apenas em inglês. Como consequência, a ampla maioria das informações contidas na Internet não está ao alcance da população com conhecimento exclusivamente de língua portuguesa.

Outros usuários amplamente beneficiados pelo uso de sistemas de RI-ML seriam os poliglotas, pois assim evitar-se-ia a necessidade da formulação da consulta em vários idiomas. O usuário poderia definir seus idiomas de domínio e um sistema de RI-ML poderia efetuar busca em todos esses idiomas utilizando uma única consulta em apenas um idioma.

Além desses, existem usuários que apesar de não dominarem um idioma estrangeiro, possuem acesso a tradutores desse idioma. Dessa forma, o usuário pode efetuar sua busca e assim identificar documentos relevantes para posterior tradução. Apesar da limitação desses usuários na formulação de consultas em idiomas estrangeiros, uma parte deles possui capacidade de leitura em outros idiomas, conseguindo assim extrair a ideia central de um texto.

1.1 Hipótese

Algoritmos de mineração de Regras de Associação (RAs) permitem identificar, de forma eficiente e rápida, a coexistência estável de um subconjunto dentro de um conjunto contido em um universo. Considerando o conjunto universo *U* como um corpus paralelo[2], e cada frase do texto como um subconjunto de *U*, é lançada a hipótese de que um algoritmo de RAs associe um termo em um idioma à sua tradução em outro idioma, portanto podendo ser usado como parte de um método para RI-ML.

1.2 Objetivos

O objetivo deste trabalho é desenvolver uma técnica para RI-ML baseada em Regras de Associação. A nova técnica deve atender aos seguintes requisitos:

- ser eficiente;
- ser escalável;
- extrair as equivalências entre termos a partir de um corpus paralelo;
- poder ser aplicada a diversos idiomas.

1.3 Metodologia

Para validar a hipótese, este trabalho pretende mapear as correspondências entre termos de uma coleção de treinamento bilíngue com base no algoritmo de mineração 'Apriori' (AGRAWAL; SRIKANT, 1994). Para tanto, este trabalho valer-se-á de semelhanças entre as etapas de RI-ML e mineração de dados. Os processos de mineração de dados consistem de quatro etapas por vezes executadas em ciclos. São elas: (*i*) identificação do problema, (*ii*) pré-processamento, (*iii*) extração e (*iv*) pós-processamento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996):

[1] http://www.translate-to-success.com/online-language-web-site-content.html

[2] Coleção de textos em que os mesmos documentos encontram-se em dois ou mais idiomas.

i. A identificação do problema consiste na compreensão do objetivo, bem como do domínio de aplicação no qual está inserido, permite uma melhor calibragem do algoritmo, além de servir como subsídio para as etapas posteriores.

ii. No pré-processamento são realizadas etapas de normalização de dados, como remoção de caracteres especiais, normalização em caso de valores numéricos, e outros processamentos visando à diminuição do volume de dados, e eventualmente uma pré-filtragem.

iii. A extração consiste na aplicação do algoritmo de mineração escolhido, normalmente por meio de um processo interativo, ajustando-se os parâmetros para atender às necessidades de retorno.

iv. O pós-processamento consiste em desenvolver métodos para filtragem, integração e avaliação do conhecimento extraído, bem como a eliminação de resultados não desejáveis. É comum nesta etapa a necessidade de intervenção por parte do usuário.

Inicialmente é definido o mapeamento de termos entre dois idiomas como objetivo desta mineração de dados. As demais etapas em muito se assemelham às etapas de RI-ML, pois ela também possui uma etapa de pré-processamento na qual a coleção é submetida à remoção de palavras comuns, remoção de acentuação, normalização dos caracteres para minúsculos e *stemming* (MANNING; RAGHAVAN; SCHÜTZE, 2008). A etapa de extração de padrões pode ser comparada à etapa de mapeamento entre conceitos num par de idiomas. O pós-processamento de RAs pode ser comparado à aplicação de métodos de avaliação dos resultados obtidos.

Dessa forma, este trabalho pretende utilizar diversos corpora paralelos e conjuntos de consultas para definir métricas de seleção de RAs que mapeiem um determinado termo em um idioma *A* a suas possíveis traduções em um idioma *B*, assim permitindo a tradução de uma consulta. Para atingir o objetivo geral do trabalho, uma série de etapas intermediárias deve ser executada, dentre as quais:

- determinação de valores para as métricas do algoritmo de mineração de RAs;
- definição das RAs que devem ser mantidas, ou seja, para um dado termo, quais as melhores traduções;
- escolha da melhor estratégia de geração de RAs.

Em RI, é imprescindível que todos os métodos propostos sejam avaliados utilizando coleções de teste apropriadas. Sendo assim, a técnica aqui proposta será avaliada comparando-a com os resultados de técnicas existentes. A fim de obter uma avaliação imparcial da abordagem proposta, participamos da campanha CLEF (*Cross Language Evaluation Fórum*) do ano de 2008 em duas trilhas, obtendo bons resultados em ambas.

1.4 Organização do trabalho

Esta dissertação está dividida em seis capítulos. Neste Capítulo foi identificada a área da ciência objeto da dissertação, o objetivo do trabalho, a hipótese, a metodologia a ser utilizada, e os métodos utilizados para avaliação dos resultados.

No Capítulo 2, são contextualizada a área de RI, e dentro desta é dada especial atenção à RI-ML e aos métodos de avaliação dos resultados produzidos por um SRI.

No Capítulo 3, são apresentadas campanhas de avaliação de resultados de trabalhos de RI-ML, e outros trabalhos recentes que abordam também RI-ML, os quais posteriormente no Capítulo 5 são comparados a este trabalho.

No Capítulo 4 são explicados mineração de dados e os seus métodos, em especial o algoritmo Apriori, bem como as regras de filtragem definidas para permitir o mapeamento entre pares de idiomas, enfatizando as similaridades entre RAs e RI-ML e definindo como um método de identificação de múltiplas traduções para um termo com suas respectivas probabilidades de ocorrência.

No Capítulo 5, são detalhados os experimentos realizados, a fim de verificar a abrangência e a independência de coleção da proposta. O trabalho proposto é comparado aos trabalhos referenciados no Capítulo 2, e a um sistema equivalente monolíngue.

No Capítulo 6, é detalhado o protótipo proposto. As conclusões e principais contribuições desta dissertação, assim como propostas de trabalhos futuros, são apresentadas no Capítulo 7.

2 LEVANTAMENTO BIBLIOGRÁFICO

Este capítulo apresenta o contexto no qual a dissertação está inserida: RI-ML. Inicialmente é contextualizada a área de RI e a subárea RI-ML. São identificadas as principais necessidades e desafios visando delimitar o escopo do problema tratado.

A RI, segundo (MANNING; RAGHAVAN; SCHÜTZE, 2008), é uma área da computação que lida com o armazenamento de documentos e a recuperação automática de informações a partir deles. É uma ciência que pesquisa a busca por informações em documentos, busca a partir dos documentos propriamente ditos ou busca por metadados que descrevem documentos. Os documentos podem ser desde dados estruturados, semiestruturados, não estruturados, imagens, vídeos, cadeias de DNA etc. Os dados podem estar isolados ou interligados em redes tais como a Internet.

Para utilizar os sistemas de RI o usuário tem de traduzir sua consulta em um conjunto de palavras-chave. A partir delas, o sistema de RI consulta sua coleção de documentos e divide ou ordena os resultados (conforme o modelo de RI empregado), para então retorná-los ao usuário. Em sistemas em que há ordenação de resultados, esses resultados são apresentados ao usuário tipicamente em ordem decrescente de similaridade com a consulta.

Visando melhorar os resultados, os sistemas de RI internamente armazenam cópias modificadas dos dados de onde são extraídas as informações. Essas cópias geralmente envolvem as seguintes mudanças:

Remoção de termos comuns: Consiste na remoção de palavras extremamente comuns e de pouco valor semântico. São escolhidas tipicamente entre as palavras de maior frequência (MANNING; RAGHAVAN; SCHÜTZE, 2008). Suas ocorrências passam a não serem indexadas e durante a busca essas palavras são excluídas do conjunto a ser buscado. Esse processo acarreta a diminuição significativa do tamanho do índice de busca, bem como a diminuição do custo computacional por consulta. Essa lista pode ser gerada manualmente por meio da seleção de conectores linguísticos, tais como artigos, conjunções, verbos de ligação etc., ou automaticamente por meio da seleção dos termos mais comuns em uma dada coleção. Conforme a quantidade de termos comuns selecionados, eles podem reduzir em mais de 50% as ocorrências de termos da coleção (ORENGO, 2004). Mecanismos de busca na *Web* como o *Google* utilizam a remoção de termos comuns para melhoria dos resultados de suas consultas (GOOGLE, 2007).

Radicalização ou *Stemming*: Consiste no processo de redução de uma palavra a sua raiz morfológica. A maior parte dos algoritmos de *stemming* são dependentes do idioma. Os primeiros estudos e o primeiro *stemmer* apareceram na década de sessenta, com o trabalho de Julie Beth Lovins (1968). Posteriormente, na década de oitenta, Martin Porter (1980) publicou seu *stemmer* para a língua inglesa. Atualmente existem diversos esforços para criar algoritmos para os mais diversos idiomas, como o RSLP para a língua portuguesa (ORENGO; HUYCK,

2001; ORENGO; BURIOL; COELHO, 2007) e para a língua espanhola (FIGUEROLA et al., 2001). Como característica adicional, o *stemming* diminui o número de termos distintos a serem indexados e permite uma expansão de consultas generalizando o termo a todas as suas formas variantes. Essa técnica é utilizada por diversos sistemas de RI na Internet, como, por exemplo, o *Google* (GOOGLE, 2007).

Indexação: A indexação é etapa essencial nos sistemas de RI, sendo uma das principais responsáveis pela otimização destes. A indexação é comumente realizada em dois níveis. O primeiro nível é um dicionário dos termos a serem indexados, e um apontador para a posição deste no segundo nível que fica responsável por listar os documentos que contêm o termo (MANNING; RAGHAVAN; SCHÜTZE, 2008). A Figura 2.1 demonstra à esquerda a lista de termos e à direita a lista dos documentos que possuem o termo. Vale ressaltar que variações otimizadas dessa técnica são amplamente utilizadas objetivando a diminuição do espaço de armazenamento necessário. Uma das estratégias mais simples é armazenar, em vez de uma lista de quais documentos possuem um dado termo, fazer-se uma lista de incrementos, anotando-se apenas quantas linhas devem ser acrescentadas para chegar-se à próxima linha em que o termo ocorre. Comumente, o primeiro nível do índice é representado por meio de árvores *trie* (árvores de prefixos) (FREDKIN, 1960). Essas árvores apresentam complexidade *O(m)*, onde *m* é a quantidade de caracteres do termo buscado, contra *O(log n),* onde *n* é o número de elementos da árvore, para árvores binárias balanceadas. Como desvantagem, esses índices aumentam a necessidade de recursos do sistema, em especial, memória. Várias técnicas são sugeridas para melhorar o acesso e o custo de criação e manutenção desse índice, comumente também são excluídos caracteres especiais. Entretanto, árvores *trie* são utilizadas em SRI atuais, como o *Zettair* (ZETTAIR, 2007).

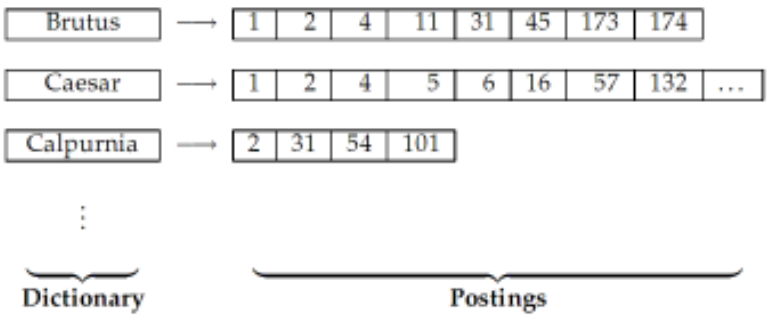

Figura 2.1: Índice invertido de busca linear (MANNING; RAGHAVAN; SCHÜTZE, 2008).

Entre outras técnicas utilizadas para a otimização de resultados em sistemas de RI, podemos citar:

Expansão de consultas: Muitas vezes os termos utilizados pelo usuário em suas consultas distinguem lexicamente dos termos disponíveis em documentos que seriam relevantes para a consulta. Esse problema se agrava ainda mais quando são formuladas consultas curtas, como as tipicamente utilizadas na Internet. Como caminho para solução desse problema propõe-se a Expansão de Consultas. A expansão pode ser implementada de diversas formas, três das mais tradicionais são a utilização de dicionários, tesauros ou busca de coocorrência de termos (XU; CROFT, 1996). A seguir as detalharemos:

- **Utilização de dicionário:** consiste na utilização de um dicionário monolíngue para buscar por sinônimos dos termos da consulta, que serão adicionados à consulta possivelmente com uma importância menor do que a do termo original.
- **Tesauros:** apesar de mais custosos do que os dicionários, permitem a generalização da consulta. Assim, além de sinônimos dos termos da busca, são adicionados termos mais gerais acerca daquele assunto.
- **Coocorrências:** procuram padrões de coocorrência entre termos que se diferenciem da distribuição padrão no corpus, por exemplo, um termo que ocorra em 0,2% do universo de documentos, entretanto apareça em 30% dos documentos recuperados pela consulta, é um termo candidato à expansão de consulta.

Realimentação de Relevantes: Muitas vezes o usuário, por desconhecer o vocabulário típico de uma dada área, pela multiplicidade dos sentidos dos termos ou pela dificuldade de sumarização de sua necessidade de consulta em termos, utiliza termos de baixa eficiência em sua consulta. Como forma de atenuar esse problema propõe-se a realimentação de relevantes, que consiste na verificação pelo usuário da relevância dos primeiros *N* resultados da consulta (onde *N* é definido em cada sistema). A partir destes, a técnica aplica uma expansão de consultas, adicionando novos termos, existentes nos documentos definidos como relevantes, os quais ainda não estavam presentes na consulta original. A técnica também ajusta os pesos dos termos da consulta com base nas avaliações de relevância fornecidas pelo usuário. Entretanto, esse processo exige a intervenção do usuário por meio da avaliação de uma lista de documentos para cada consulta. Como alternativa, então se aplica a pseudorrealimentação de relevantes, que consiste em assumir os primeiros *N* resultados como relevantes e aplicar a reformulação da consulta com esses resultados.

2.1 Modelos de Recuperação de Informações

Modelos de RI utilizam-se de diferentes grupos de premissas. Segundo Baeza-Yates e Ribeiro-Neto (1999), um modelo de RI pode ser caracterizado como uma quádrupla $[D,Q,F,R(q_i,d_j)]$, onde:

D é um grupo composto de visões lógicas para os documentos da coleção.

Q é um grupo composto de visões lógicas das necessidades de informação dos usuários (consultas).

F é um *framework* para modelar a representação dos documentos, consultas e suas relações.

$R(q_i,d_j)$ é uma função de ordenação que associa um número real com a consulta $q_i \in Q$ e a representação do documento $d_j \in D$, associando, assim, uma ordenação por relevância (similaridade entre o documento e a consulta).

Vários modelos de RI foram propostos. Como modelos clássicos são referenciados na literatura os modelos Booleano, Vetorial (SALTON; WONG; YANG, 1975) e Probabilístico (ROBERTSON; SPARCK-JONES, 1976). A partir desses modelos surgiram outros, como o baseado em Lógica Nebulosa (OGAWA; MORITA; KOBAYASHI, 1991), o Modelo de Espaço Vetorial Generalizado (MANNING; RAGHAVAN; SCHÜTZE, 2008), Indexação

Semântica Latente (DEERWESTER et al., 1990), e Redes Neurais (ROSS; PHILIP, 1991). A seguir são explicados os modelos clássicos.

2.1.1 Modelo Booleano

Trata-se do mais simples de todos os modelos de RI. Baseia-se na partição do conjunto *D* segundo os termos *Q*. Para essa partição, utilizam-se operadores lógicos **AND**, **OR** ou **NOT**, podendo estes serem explicitados na consulta (BAEZA-YATES; RIBEIRO-NETO, 1999), ou deixados conforme a lógica interna do Sistema de Recuperação de Informações (SRI). Para essa abordagem, os documentos são tratados como pacotes de palavras, nos quais a ordem destas é irrelevante. Na Figura 2.22 é demonstrada uma consulta em que se busca documentos com o termo *kb* e que não possuam os termos *ka* e *kc*.

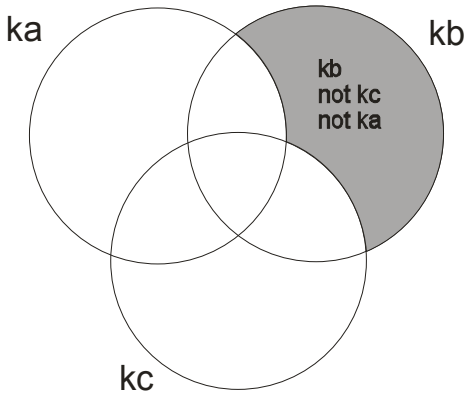

Figura 2.2: Diagrama de Vens como representação gráfica do modelo Booleano.

A maior desvantagem desse modelo é sua incapacidade de gerar uma ordenação dos resultados obtidos, não atribuindo um grau de similaridade a cada documento recuperado em relação à consulta proposta. Mesmo assim, o modelo é usado largamente em RI, devido ao seu baixo custo computacional e à simples implementação. Outro problema desse modelo é que muitas necessidades de informações não são facilmente transformadas em consultas Booleanas devido a sua complexidade (MANNING; RAGHAVAN; SCHÜTZE, 2008). Derivações desse modelo, como o modelo booleano estendido, minimizam esse problema.

2.1.2 Modelo vetorial

O Modelo Vetorial, inicialmente proposto por Salton (1971), utiliza pesos para minimizar as limitações do modelo Booleano. O modelo utiliza métricas para definir o peso de cada termo em cada documento, bem como o grau de similaridade entre cada um dos documentos e a consulta (BAEZA-YATES; RIBEIRO-NETO, 1999).

Dentro desse modelo, cada documento é representado como um vetor em um espaço de *n* dimensões onde *n* é o total de termos da coleção. Para identificar a similaridade entre um dado documento *d* e a consulta *q*, é proposto o cálculo do cosseno (Equação 2.1) entre seus vetores, sendo:

$$sim(d_j, q) = \frac{(\vec{d}_j \bullet \vec{q})}{|\vec{d}_j| \times |\vec{q}|}$$

Equação 2.1: Similaridade modelo vetorial.

A similaridade é definida como a distância entre o vetor representante da consulta e os vetores representantes de cada um dos documentos. Isso torna possível uma ordenação total dos resultados, permitindo assim um efetivo ranqueamento. Outra importante característica é que esse modelo permite resultados aproximados. Como desvantagens, esse modelo não leva em consideração a dependência entre os termos (BAEZA-YATES; RIBEIRO-NETO, 1999; MANNING; RAGHAVAN; SCHÜTZE, 2008).

Como forma de não beneficiar demasiadamente termos comuns na coleção ou documentos longos, o modelo vetorial utiliza-se da normalização dos vetores, evitando assim que documentos demasiadamente longos sejam beneficiados; além disso, esse modelo também pondera a importância de cada termo de uma consulta. Para que isso possa ocorrer, o modelo utiliza-se de dois cálculos intermediários, o cálculo da frequência de um termo (TF), e da frequência inversa em documentos (IDF). A partir dessas duas métricas é possível definir a importância de um termo em um dado documento, conforme a Equação 2.2:

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Equação 2.2: Cálculo de frequência do termo no documento.

Onde o numerador é o número de ocorrências do termo considerado no documento d_j, e o denominador o número de ocorrências de todos os termos no documento d_j.

O IDF (*Inverse document frequency*) Equação 2.3 é uma medida da importância geral do termo (obtido dividindo o número de todos os documentos N pelo número de documentos contendo o termo n_i e, em seguida, tomando o logaritmo desse quociente) (WIKIPÉDIA, 2008).

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

Equação 2.3: Cálculo de frequência inversa de termo.

2.1.3 Modelo probabilístico

Proposto inicialmente por Robertson e Sparck-Jones (1976 apud ORENGO, 2004), é também reconhecido como *binary independency model*. Esse modelo baseia-se em pesos binários que representam a presença ou ausência de termos nos documentos. O conjunto desses pesos resulta em um vetor que identifica a probabilidade de um documento ser relevante para a consulta ou não. Esse modelo utiliza-se do princípio da ordenação baseado no teorema de Bayes.

O Modelo Probabilístico considera um processo interativo visando convergir para uma similaridade estimada. Assim, para calcular $P(+Q|d)$ onde d é um documento e q uma consulta e $P(-Q|d)$ a probabilidade de um documento, d não ser relevante para a consulta q. Assim, o documento é considerado relevante se $P(+Q|d) > P(-Q|d)$ o vetor $W_{d|q}$ é calculado através da fórmula enunciada na Equação 2.4.

$$w_{d|q} = \frac{P(+Q_q \mid D)}{P(-Q_q \mid D)}$$

Equação 2.4: Similaridade documento consulta no modelo probabilístico.

Segundo Cardoso (CARDOSO, 2000), além do bom desempenho prático, o princípio probabilístico de ordenação, uma vez garantido, resulta em um comportamento ótimo do método. Entretanto, a desvantagem é que esse comportamento depende da precisão das estimativas de probabilidade. Além disso, o método não explora a frequência do termo no documento.

No início do processo, imediatamente após a entrada da consulta, não temos informação sobre documentos relevantes. Entretanto, como precisamos de dados iniciais para as convergências, adota-se comumente as seguintes suposições: assume-se uma relevância inicial igual para todos os termos indexados, tipicamente 0,5 e que a distribuição dos termos indexados é executada de forma aleatória. A partir dessa suposição são feitas induções utilizando as fórmulas da Equação 2.5:

$$P(k_i \mid rel) = 0.5$$
$$P(k_i \mid nrel) = \frac{n_i}{N}$$

Equação 2.5: Probabilidades iniciais do modelo probabilístico.

Onde a probabilidade de o termo indexado k_i estar presente em um documento relevante é $P(k_i \mid rel)$ e a probabilidade de um de um termo indexado estar presente em um documento irrelevante é $P(k_i \mid nrel)$, e n_i é o número de documentos que contêm o termo indexado e N é o número total de documentos da coleção.

Dentre os diversos modelos probabilísticos de RI, destaca-se o Okapi BM25. Este modelo é muito utilizado e com freqüência serve de *baseline* em experimentos.

$$\text{BM25}(D,Q) = \sum_{i=1}^{n} idf(q_i) \cdot \frac{tf \cdot (k_1 + 1)}{tf + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avg}})}$$

Equação 2.6: Modelo probabilístico BM25

onde b um valor de normalização do tamanho dos documentos. Valores b próximos a 0 fazem com que documentos maiores tenham seus termos com pesos próximos ao que teriam em documentos mais curtos. O valor 1 faz com que o somatório dos pesos dos termos em documentos de quaisquer tamanhos sejam iguais. A variável k define a importância da presença de cada termo na consulta. Um valor k=0 corresponde a um modelo binário em que é atribuído peso 1 quando o termo esta presente e 0 quando não está. Valores maiores de k, ponderam a similaridade em função da frequência do termo no documento. A variável $|D|$

corresponde ao tamanho do documento *D* e a variável *avg* ao tamanho médio dos documentos da coleção.

2.1.4 Avaliação em recuperação de informações

A forma mais intuitiva de avaliar a qualidade de resposta de um SRI seria a simples avaliação de quantos resultados relevantes ele retornou. Porém, algumas etapas devem ser formalizadas para permitir uma melhor comparação entre sistemas. Não seria útil um sistema retornar diversos documentos relevantes e todos estarem esparsamente distribuídos na lista de documentos recuperados, bem como também não seria justo considerar iguais dois sistemas que para uma dada consulta resultam ambos em cinco resultados relevantes, no entanto o primeiro de um total de oito retornados e o segundo de um total de trinta retornados. Assim, como forma de padronizar a avaliação de sistemas de RI, foram desenvolvidas algumas métricas. A Figura 2.3 exemplifica as diferenças entre o conjunto dos documentos recuperados e o conjunto dos documentos relevantes para uma dada consulta. A intersecção de ambos os conjuntos representa os documentos corretamente recuperados. Porém, como é visto na Figura, existem documentos relevantes não recuperados, bem como documentos irrelevantes recuperados.

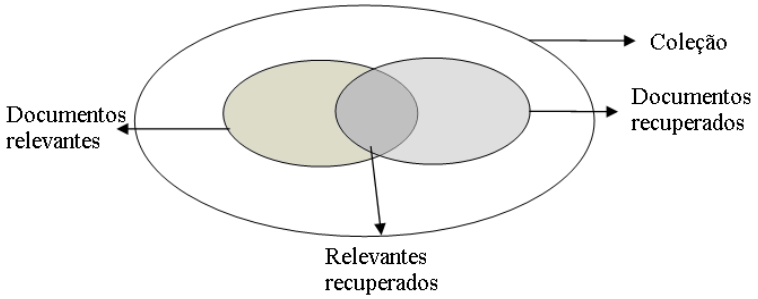

Figura 2.3: Precisão e revocação.

É importante ressaltar que com o objetivo de avaliar SRI, o julgamento de relevância é feito de forma binária, assim sendo, um documento é relevante ou irrelevante, não havendo ordenação de mais ou menos relevante.

Com base nesses dois conjuntos de documentos relevantes e documentos recuperados, foram elaboradas duas métricas tradicionais para avaliação de SRI:

Revocação: A fração dos documentos relevantes que foi recuperada conforme Equação 2.:

$$\text{Revocação} = \frac{\text{Total de relevantes recuperados}}{\text{Total de relevantes}}$$

Equação 2.7: Revocação.

Precisão: A fração dos documentos recuperados que são relevantes conforme Equação 2.8:

$$\text{Precisão} = \frac{\text{Total de relevantes recuperados}}{\text{Total de recuperados}}$$

Equação 2.8: Precisão.

Supondo que para uma dada consulta são retornados 8 documentos, e destes 3 são relevantes de um total de 5 relevantes existentes em uma coleção, então, a revocação para essa consulta será 60% (revocação=3/5) e a precisão será 37,5% (precisão=3/8). Altos índices de revocação são comumente acompanhados de baixos índices de precisão e vice-versa.

É desejável ter apenas um valor que sumarize precisão e revocação. Uma alternativa para isso é a combinação de precisão e revocação em uma única medida chamada *F-measure.* Essa medida possui também duas importantes propriedades: penaliza sistemas com alta precisão e baixa revocação e vice-versa, além de possuir um descriminante β, que permite enfatizar tanto precisão quanto revocação. Para manter pesos equivalentes para precisão e revocação, utiliza-se $\beta = 1$, para enfatizar revocação N vezes utiliza-se $\beta = N$, já para enfatizar N vezes precisão, utiliza-se $\beta = 1/N$.

Porém, essas métricas apresentam uma limitação, pois não consideram a posição do documento retornado. Assim, no exemplo anterior, os documentos retornados poderiam tanto estar nas posições sexta, sétima e oitava do *ranking*, ou nas três primeiras, que o resultado em termos de revocação, precisão ou *F-measure* seria o mesmo. Objetivando premiar sistemas que retornam elementos relevantes no topo do *ranking,* foi desenvolvida a medição da precisão em 11 pontos de revocação, retornando ao exemplo anteriormente citado e supondo a lista de documentos retornada abaixo, na qual em negrito encontram-se os documentos relevantes.

Tabela 2.1: Exemplo de *ranking*.

D1	D5
D2	D6
D3	D7
D4	**D8**

Como existem 5 documentos relevantes na coleção, cada documento relevante corresponde a 20% do total de relevantes. Para medição dos 11 pontos de revocação, são percorridos todos os relevantes retornados, e para fins de cálculo, no primeiro ponto se tem 1 relevante de 1 retornado 100% de precisão, para o segundo relevante D3, têm-se 2 relevantes retornados de 3 documentos retornados 66,66%, para o relevante seguinte D8, têm-se 3 relevantes retornados de 8 documentos retornados 37,5%.

A partir desses valores deve-se executar uma interpolação para os 11 pontos revocação. Para cada um dos pontos de revocação de 0 a 1, com passo de 0,1, é visto qual a maior precisão para uma revocação maior ou igual a atual. Assim, para o exemplo acima, a curva de precisão *versus* revocação ficaria conforme a Figura 2.4.

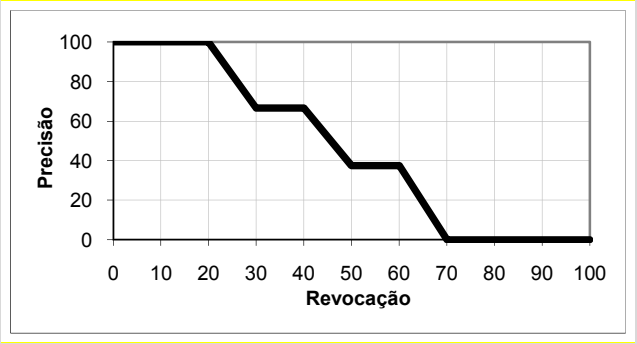

Figura 2.4: Gráfico precisão x revocação.

Esse gráfico apresenta a precisão *versus* revocação para uma consulta, mas usualmente em um SRI são executadas diversas consultas a fim de avaliar o sistema. Assim, para gerar um gráfico que sumarize as informações de várias consultas, se calcula a média das precisões em cada ponto dessa curva, e por final se "monotoniza" a curva.

Outra forma de sumarizar a informação é por meio da precisão média (*Average Precision - AP*), que se refere às médias não interpoladas das precisões médias para uma consulta. A AP é computada para cada documento relevante recuperado, utilizando precisão zero para os documentos relevantes que não foram retornados. No exemplo acima, AP seria 100% referente ao primeiro documento recuperado acrescido de 66% referente ao segundo documento recuperado e de 37,5% referente ao terceiro recuperado, a isso adiciona-se duas vezes 0% referente aos dois documentos não recuperados. A média desses valores 40,83% é a AP da consulta. Para um conjunto de consultas, calcula-se a média aritmética de suas APs, obtendo-se a **MAP** (*mean average precision*), que é amplamente utilizada em campanhas de avaliação, como o *Text Retrieval Conference* (TREC)[3] ou o *Cross-Language Evaluation Forum* (CLEF)[4], pois sumariza em um único número a precisão de um dado sistema para um conjunto de consultas.

2.2 Recuperação de Informações Multilíngues

Recuperação de Informações Multilíngues (RI-ML) consiste na busca de documentos em um idioma a partir de uma consulta formulada em outro idioma. Esse cruzamento linguístico distingue a RI-ML da Recuperação de Informações Tradicional (GEY; KANDO; PETERS, 2005). As primeiras pesquisas sobre RI-ML foram propostas por Salton (1971), que demonstrou que utilizando um tesauro construído manualmente, a RI-ML pode ter resultados próximos aos da RI tradicional (monolíngue). Para tanto, a proposta era o agrupamento de palavras em classes, em cada um dos dois idiomas utilizados nos testes (Inglês e Alemão), assim servindo como um *link* entre termos de idiomas diferentes. Durante a indexação, a proposta localiza no tesauro a classe de cada um dos termos e então troca esse termo pelo indexador da classe. Entretanto, nem sempre era possível atribuir uma classe ao termo da consulta, por exemplo, em nomes próprios.

[3] THE TEXT RETRIEVAL CONFERENCE (TREC). Desenvolvido por TREC. Disponível em: <http://trec.nist.gov>. Acesso em: 11 out. 2007.

[4] CROSS-LANGUAGE EVALUATION FORUM (CLEF). Desenvolvido por CLEF. Disponível em: <http://www.clef-campaign.org>. Acesso em: 18 set. 2007.

Dentre os problemas de RI-ML, Grefenstette (1998) define três principais:

- O primeiro problema é saber como é que um termo expresso em um idioma pode ser escrito em um outro idioma. Isso tem a ver com alternativas para atravessar a barreira do idioma.
- O segundo problema é decidir quais das possíveis traduções deverão ser mantidas. Manter mais de uma tradução é útil na promoção da revocação. No entanto, usando traduções inadequadas, reduzirá a precisão.
- O terceiro problema é decidir como fazer corretamente a pesagem de importância das traduções alternativas quando mais de uma é selecionada.

Segundo Gey, Kando e Peters (2005), a pesquisa nessa área apresenta três desafios, além do problema principal que é o mapeamento de conceitos:

- Obter recursos para algumas línguas: consiste na dificuldade em conseguir corpora paralelos, tesauros ou até mesmo tradutores para línguas não oficiais das Nações Unidas, e principalmente para dialetos.
- O fato de não haver corpus multilíngue na Internet de grande tamanho e o custo de criação e manutenção de um corpus seja paralelo ou comparável, que não esteja entre os 10 idiomas mais falados no mundo e a dificuldade de buscas em idiomas que possuem mais de uma representação lexicográfica, como o japonês e o chinês.
- A dificuldade de adaptação dos motores de busca, principalmente para línguas que não usam alfabeto latino.

A seguir são demonstradas abordagens utilizadas em RI-ML e posteriormente alguns trabalhos de RI-ML, os quais foram avaliados em comparação com a proposta desta dissertação.

Objetivando solucionar o problema da RI-ML, diversas soluções foram propostas. As abordagens podem ser divididas em quatro grandes grupos, conforme a estratégia para o mapeamento de conceitos utilizada:

- **Tradução automática (TA):** essa é a estratégia mais direta de RI-ML, consiste em traduzir automaticamente a consulta para o idioma dos documentos e prosseguir com um processo de RI tradicional. O problema dessa abordagem é que as consultas geralmente são compostas por poucas palavras soltas, o que não fornece contexto para que os sistemas de TA selecionem a tradução adequada para cada termo (note que apenas uma tradução é selecionada). Se uma tradução equivocada for escolhida, documentos irrelevantes serão recuperados. Além disso, nem sempre está disponível um *software* de TA para os idiomas que se deseja executar consultas em RI-ML. Com essa abordagem podemos citar Fujii e Ishikawa (2004).
- **Tesauro:** nessa estratégia, o usuário informa os termos da consulta e o sistema os procura em um tesauro multilíngue, fazendo a substituição dos termos originais pelos sinônimos encontrados no tesauro. Experimentos com tesauros geralmente foram feitos com "vocabulário controlado", ou seja, coleções em que palavras do tesauro são atribuídas aos documentos como descritores. Essa abordagem geralmente atinge bons resultados. Por outro lado, existe o alto custo de construção e manutenção de um tesauro. Entre os trabalhos baseados tesauros multilíngues, podemos citar Gey e Jiang (1999).

- **Dicionário eletrônico:** essa abordagem funciona de maneira semelhante à anterior. Nela os termos da consulta são substituídos pelas traduções encontradas no dicionário. O maior incentivo a essa abordagem é a crescente disponibilização de dicionários eletrônicos. A maior desvantagem é que esses dicionários são desenvolvidos para o uso humano, o que dificulta o seu processamento pelos SRIs. Por exemplo, formas variantes das palavras (como femininos, plurais, aumentativos) não estão normalmente presentes nos dicionários. Além desses problemas, podemos citar a contextualização da tradução e a ausência de entidades nomeadas.
- **Baseadas em Corpus:** as abordagens baseadas em corpus analisam coleções de textos em vários idiomas. Essas coleções podem ser paralelas (os mesmos documentos em dois ou mais idiomas) ou comparáveis (os documentos não são os mesmos, mas tratam do mesmo assunto). O objetivo é extrair de forma automática a informação necessária para saber como um termo (ou conceito) pode ser mapeado para outro idioma. O problema dessa abordagem é a falta de corpora. Contudo, atualmente uma solução que vem sendo cada vez mais utilizada é a geração de corpus paralelo/comparável através de mineração de documentos na *Web*. Como abordagens baseadas em corpus, podemos citar Deerwester et al. (1990) e Orengo e Huyck (2003).

Os trabalhos recentes muitas vezes mesclam diferentes técnicas de forma a obter um melhor resultado. Além disso, a aplicação das técnicas de RI-ML pode ser executada sobre as consultas ou sobre a coleção, sendo o primeiro alvo a opção mais pesquisada, pela baixa escalabilidade da aplicação dos métodos de RI-ML no corpus inteiro (SANDERSON, 1994).

3 TRABALHOS RELACIONADOS

Dentro do contexto de RI-ML, são descritas as principais campanhas de avaliação de SRI, dada a importância da mensuração dos resultados para a área. Também são descritos alguns trabalhos relacionados que foram desenvolvidos recentemente, juntamente com seus resultados. Entre os trabalhos relacionados, dá-se maior enfoque aos que participaram do CLEF, por serem passíveis de comparação direta e idônea, já que a proposta posteriormente apresentada neste trabalho também participou dessa avaliação.

3.1 Campanhas de avaliação

Existem campanhas dedicadas exclusivamente à avaliação de SRI, das quais as mais importantes são o *Cross Language Evaluation Fórum* (CLEF), o *NII Test Collection for IR* (NTCIR, 2008) e o *Text REtrieval Conference* (TREC, 2008). Nessas campanhas há uma avaliação idônea do desempenho das propostas a partir de consultas de documentos providos pelos organizadores. A avaliação dos resultados também é feita pelos organizadores da campanha e não pelos próprios participantes. A seguir é detalhado o funcionamento dessas campanhas.

3.1.1 TREC

Originalmente destinada à avaliação de SRI em língua inglesa, o TREC se expandiu para outros idiomas a partir da sua terceira edição. Inicialmente foram incluídos tópicos em espanhol e posteriormente tópicos em chinês. A partir da oitava edição, passaram a ser aceitas também outras línguas europeias, como alemão, francês e italiano. Com o surgimento do CLEF, o TREC passou a não tratar mais das línguas europeias. Em seguida, com a criação do NTCIR, o TREC deixou de incluir chinês, japonês e coreano. Nos anos seguintes, apenas alguns trabalhos envolvendo árabe foram apresentados na conferência, entretanto sem uma sessão específica (GEY; KANDO; PETERS, 2005).

Uma de suas principais contribuições para á área de RI, e por consequência, para RI-ML, é seu modelo de avaliação, o qual acabou por ser referência e adotado por diversas outras campanhas. Os SRIs participantes são avaliados em coleções de teste. Uma coleção de teste é composta por documentos, consultas e julgamentos de relevância. A Figura 3.1 apresenta um exemplo de tópico de consulta. Os participantes recebem a coleção de documentos e os tópicos de consulta e executam seus métodos de RI sobre eles. Como resultado, cada participante envia a lista de documentos recuperados em resposta a cada consulta. Com base nos resultados enviados pelos participantes, os organizadores da campanha realizam os julgamentos de relevância, ou seja, definem o conjunto de documentos relevantes para cada tópico de consulta. O julgamento de relevância acontece segundo um método de *Pooling* proposto por Sparck-Jones e Rijsbergen (1975), que consiste na avaliação de relevância binária dos primeiros N registros enviados por cada grupo da campanha de avaliação, qualquer documento que não conste dentre os avaliados é considerado irrelevante. Por fim, as

avaliações dos sistemas participantes, juntamente com os artigos que descrevem os experimentos executados, são publicadas em anais.

No caso de RI-ML, os tópicos de consulta são traduzidos manualmente por pessoas que possuam fluência em ambos os idiomas.

```
<top>
<num> Number: 307
<title> New Hydroelectric Projects
<desc> Description:
Identify hydroelectric projects proposed or under construction by
country and location. Detailed description of nature, extent,
purpose, problems, and consequences is desirable.
<narr> Narrative:
Relevant documents would contain as a minimum a clear statement
that a hydroelectric project is planned or construction is under
way and the location of the project. Renovation of existing
facilities would be judged not relevant unless plans call for a
significant increase in acre-feet or reservoir or a marked change
in the environmental impact of the project. Arguments for and
against proposed projects are relevant as long as they are supported
by specifics, including as a minimum the name or location of the
project. A statement that an individual or organization is for or
against such projects in general would not be relevant. Proposals
or projects underway to dismantle existing facilities or drain
existing reservoirs are not relevant, nor are articles reporting a
decision to drop a proposed plan.
</top>
```

Figura 3.1: Exemplo de tópico TREC.

3.1.2 NTCIR

O (*National Institute of Informatics*) *Test Collection for IR* (NTCIR) é uma campanha de avaliação que ocorre a cada 18 meses na Ásia. Seu foco é o acesso à informação, que inclui além de RI e RI-ML, também sumarização de texto e mineração de dados. Essa campanha caracteriza-se por ter seções de RI-ML entre línguas muito diferentes sintaticamente, como japonês-inglês e chinês-inglês. O NTCIR adota um sistema de avaliação de resultados semelhante ao do TREC, modificando apenas as coleções de dados (GEY; KANDO; PETERS, 2005).

3.1.3 CLEF

Trata-se de uma campanha de avaliação realizada anualmente desde 2000 na Europa. O CLEF está em sua décima edição e possui grande importância, dado o grande número de idiomas da União Europeia. Tem como objetivos: avaliar métodos de RI e RI-ML, incentivar a transformação de sistemas RI para RI-ML e concentrar discussão sobre RI e RI-ML.

O processo de avaliação é idêntico ao do TREC para as trilhas de RI e RI-ML. No entanto, o corpus distribuído, bem como o número de consultas, varia ao longo dos anos. Em suas primeiras edições o corpus era de notícias de jornais e, em 2008, de metadados bibliográficos.

Atualmente a campanha se expandiu para várias áreas de RI, como busca por imagens, busca em *blogs*, em fotos, vídeos, informações geográficas, entre outras. Ao longo de suas nove edições, já houve trabalhos de RI-ML em alemão, espanhol, francês, finlandês, inglês, italiano, português, russo e sueco, entre outros.

3.2 Trabalhos recentes

O CLEF, durante suas primeiras oito edições, baseou-se exclusivamente em coleções de notícias de jornais. Entretanto, em 2008, em sua trilha tradicional de RI-ML, a coleção era composta por metadados bibliográficos de bibliotecas digitais. A Figura 3.2 demonstra uma referência bibliográfica. Essa mudança adicionou novos desafios e problemas. A principal vantagem foi que todos os grupos não possuíam conhecimento do corpus até recebê-lo, e técnicas preparadas exclusivamente para os corpora antigos falhariam. Esse corpus adicionou novos desafios, os dois principais são: o corpus deixa de ter estruturação dissertativa, passando a ter estrutura de banco de dados exportado em XML, com campos nomeados e *tipados*. Além disso, passou a ter alguns campos em idiomas diferentes ao da coleção, por exemplo, títulos de livros em idiomas estrangeiros, ou títulos originais de obras traduzidas. Outros problemas que podem ser citados são a não existência de todos os campos em todos os registros e o tamanho pequeno dos documentos.

Para essa trilha foram preparadas 50 consultas em cada um dos idiomas (alemão, francês, holandês e espanhol, a partir das consultas originalmente em inglês), as quais são aplicadas sobre um corpus de 1.000.100 registros, e pelo menos 2533 documentos relevantes.

```
<document>
<DC>
<title>Our parish : the story of St. Vincent de Paul's, Clapham Common,
London.</title>
<language>eng</language>
<description>"This supplement should be read as the opening chapters of the
book itself" _ 1st leaf.</description>
<description>Previous ed.: 197-?.</description>
<description>Author: Lawrence P. Seglias.</description>
<subject>282/.42166</subject>
<subject>St. Vincent de Paul's (Church : Clapham Common)</subject>
<subject>London Wandsworth (London Borough) Clapham Catholic Church St Vincent
de Paul's (Church : Clapham Common)</subject>
<identifier></identifier>
<type>text</type>
<identifier>011185963.</identifier>
<location>British Library HMNTS X.809/61406</location>
</DC>
</document>
```

Figura 3.2: Exemplo de documento do CLEF 2008.

Para a trilha de desambiguação semântica, mantiveram-se as mesmas coleções dos anos anteriores. Para a língua inglesa, as coleções eram formadas por um ano do *Los Angeles Times* de 1994 e do *Glasgow Herald* de 1995. Nessa trilha, os termos foram anotados com suas classes gramaticais.

```
<DOC>
<DOCNO>GH950321-000000</DOCNO>
<DOCID>GH950321-000000</DOCID>
<DATE>950321</DATE>
<HEADLINE>Siege gunman is jailed for 14 years</HEADLINE>
<EDITION>1</EDITION>
<PAGE>4</PAGE>
<RECORDNO>979288175</RECORDNO>
<TEXT>
A GUNMAN was jailed for 14 years yesterday following a siege in which he held
a sawn-off shotgun in the mouth of a woman hostage.
The High Court in Perth was told that, following her ordeal, Ms Ann Hutcheon,
27, suffered from severe post-traumatic stress disorder and has been under
psychiatric care in hospital for the past three months.
Judge Lord Gill told Wayne McGettigan, 25, that Miss Hutcheon and four
policemen were fortunate not to have been murdered during the seige.
Information that McGettigan had a sawn-off shotgun led the police to surround
his lodgings in Christies Lane, Montrose, last November 27.
McGettigan admitted he held armed officers at bay, he presented the gun at two
of his guests, Miss Hutcheon and her boyfriend Mr Stuart Buist, detained them
against their will, put the gun at the woman's head and in her mouth, and
compelled her to shield him from police marksmen as he moved to another house
in the same street occupied by his mother.
He also admitted that, during the siege, he fired the gun twice at a total of
four policemen.
McGettigan eventually released Miss Hutcheon as he retreated into his mother's
home where he took sleeping tablets He was arrested in a semi-conscious state
four hours after the siege began.
</TEXT>
</DOC>
```

Figura 3.3: Exemplo documento CLEF.

A seguir são detalhados os trabalhos recentemente publicados no CLEF que envolvam RI-ML, bem como alguns artigos recentes publicados no SIGIR[5] acerca desse tema.

3.2.1 *Xtrieval*

Algoritmo desenvolvido pela Universidade Tecnológica de Chemnitz (Alemanha). Baseia-se no motor de busca *extensible retrieval and evaluation* (KÜRSTEN; WILHELM, T.; EIBL, 2008), o qual é baseado no motor *Lucene* para RI. O *Lucene* é um motor de busca escrito em Java e portado para diversas linguagens. É desenvolvido pela fundação Apache. O *Xtrieval* difere do *Lucene* por aplicar a fusão de dados, sobre o método *Z-score*.

A fusão de dados consiste na junção de dois ou mais métodos de atribuição de similaridade entre documentos e consultas. Neste trabalho mesclou-se os métodos *Okapi* (ROBERTSON et al., 1994) e *LNu-Ite*, formando assim o *Z-score*. Considerando RSV_k, o índice de similaridade entre um dado documento e uma consulta; $Mean_i$, o valor médio de *RSV* para qualquer termo do corpus no motor *i*; $Stdev_i$, o desvio padrão de *RSV* no motor *i*; e α_i, o peso relativo dado ao motor *i* (SAVOY, 2005), o modelo *Xtrieval* define a similaridade entre um documento e uma consulta conforme a Equação 3.1:

$$\text{Z-Score}_k = \sum_{i=0}^{i=1} \alpha_i \times [((RSV_k - Mean_i) / Stdev_i) + ((Mean_i) - Min_i) / Stdev_i)]$$

Equação 3.1: Cálculo do *Z-Score*.

[5] SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL (SIGIR). Desenvolvido por SIGIR. Disponível em: <http://www.sigir.org/>. Acesso em: 15 jun. 2008.

Essa fórmula objetiva a ponderação entre os dois métodos de RI utilizados no trabalho, pois os escores absolutos deles não são normalizados.

Para o mapeamento de conceitos em RI-ML, utiliza-se um TA, o *Google* tradutor. O trabalho também utiliza realimentação de pseudorrelevantes considerando os primeiros 10 resultados relevantes, e o removedor de sufixos Porter (1980).

Com esse método, a proposta apresentou no CLEF 2008 resultados entre 87,81 e 95,9% dos obtidos nas consultas monolíngues, variando em função do idioma destino escolhido. O melhor resultado foi obtido na trilha de RI-ML com corpus em inglês. Entretanto, o método proposto depende da existência de TA entre os idiomas envolvidos, além de ser custoso computacionalmente, tendo demorado 6 horas para serem executadas as 50 consultas propostas.

3.2.2 XRCE

Utiliza um tesauro monolíngue para expansão de consultas (ver Capítulo 2). Para a identificação dos idiomas dos campos dos documentos utiliza-se dicionários bilíngues gerados a partir de um corpus paralelo artificial e dicionários monolíngues para identificação de pertinência de termo a um dado idioma. Além de casamento de *strings* por similaridade em substituição a remoção de sufixos (que é usado apenas na língua Alemã), utiliza um decompositor de palavras desenvolvido para o XRCE (CLINCHANT; RENDERS, 2009).

O algoritmo considera a existência de *N* idiomas, no caso proposto, três (inglês, francês e alemão), definindo o idioma do documento como sendo o idioma mais provável, recebendo uma probabilidade α (ex. 0,8). As duas outras línguas recebem uma probabilidade de *(1-α)/2*. Com a definição desses valores, executa-se o mapeamento do termo no dicionário. Além disso, são buscados termos mais gerais em um tesauro. Esse conjunto de traduções passa a ser utilizado na consulta em substituição à tradução do termo original. Caso a palavra venha a constar de uma lista de *stopwords*, ela será ignorada nesta etapa. Após a identificação dos termos e seu mapeamento para a língua do corpus, a consulta é executada através do motor de RI *Lemur* (LEMUR, 2009). A fim de melhorar sua performance, é proposta a adaptação de dicionário que permite o mapeamento de mais de um termo destino para o mesmo termo origem, proposta pelo algoritmo a *cross-entropy* (Equação 3.2).

$$ce(q_s \mid d_t) = \sum_{w_t, w_s} P(w_t \mid w_s) \times P(w_s \mid q_s) \times \log P(w_t \mid d_t)$$

Equação 3.2: *Cross-Entropy*.

Onde a consulta original é formada por qs=$(w_{s1},....,w_{sm})$, e $P(w_s|q_s)$ é a relevância do documento q_s para o termo w_s baseado no modelo probabilístico, analogamente para w_t, e $P(w_t|w_s)$ é a similaridade entre dois termos. A partir do *ranking* gerado após atribuição de relevância para todo d_t, são selecionados como pseudorrelevantes os 50 documentos melhor ranqueados, aplicado novamente ao motor de RI *Lemur*.

Essa técnica apresenta algumas inconsistências: considera a existência de apenas um trio de idiomas, depende de uma série de recursos para o mapeamento da consulta entre idiomas, incluindo tesauros, dicionários monolíngues, corpora paralelos e a necessidade de geração de um dicionário bilíngue, o que eleva o custo computacional.

Os resultados obtidos foram MAP de 28,25% para a versão bilíngue e 34,66% para a monolíngue, o que resulta em um desempenho de 80,5% do resultado monolíngue. Entretanto,

essa diferença foi maior em outros idiomas como, por exemplo, na versão bilíngue para francês, em que alcançou 46,57% do resultado da versão monolíngue.

3.2.3 *WikiTranslate*

Trata-se de uma proposta que se baseia na utilização da Wikipédia como um corpus paralelo comparável. Essa proposta apresentaria grandes vantagens, como o tamanho do vocabulário em constante crescimento pelo trabalho comunitário dos usuários, e a categorização dos verbetes com granularidade mais fina que um dicionário, o que permite uma melhor desambiguação. A presença de *links* para o termo, assunto e sua tradução em outros idiomas (com marcador semântico específico) além da aglutinação de sinônimos, abreviações e variações regionais sobre o mesmo verbete, permitem ao algoritmo definir traduções e contexto do artigo. Por outro lado, a Wikipédia não possui verbetes para palavras comuns (ter, estar etc.).

A proposta se divide em dois passos. Primeiramente é preciso mapear os conceitos da consulta na Wikipédia. Isso pode ser feito de duas formas: através dos *links* internos, ou utilizando a parte do corpo do artigo. O passo seguinte consiste no mapeamento da consulta para o idioma destino. Nesse passo podem ser mapeados mais de um conceito por meio da expansão de consultas. Para isso utiliza-se os redirecionamentos que chegam e que partem de cada artigo da Wikipédia, assim, mapeando também sinônimos e variantes lexicográficas regionais dos termos (NGUYEN et al., 2009).

A técnica, apesar de simples, apresenta alguns problemas: a não utilização das categorizações disponíveis na Wikipédia, o que poderia ser considerado um pequeno tesauro; e a não exploração de forma consistente dos *links* de *cross-language*, assim identificados, que possuem o mapeamento para o termo em outro idioma. Os autores justificam que estes podem levar a seções de artigos em outros idiomas e não a artigos completos. Entretanto, seria simples definir se trata-se de seção ou não. Por fim, o descarte de *links* de desambiguação poderia ser mais bem explorado, por meio da busca apenas no subconjunto resultante da desambiguação.

Os resultados obtidos foram de MAP 34,07% para a monolíngues e MAP de 22,78% para a versão bilíngue, o que representa 66,86% do resultado monolíngue. Entretanto, essa diferença foi maior em outros idiomas como, por exemplo, na versão bilíngue com consultas em alemão, na qual o resultado alcançou 59,82% do resultado obtido pela versão monolíngue.

3.2.4 *Logistic Regression*

Esse algoritmo é constantemente avaliado pelo CLEF. Ele se baseia em um modelo probabilístico e em coeficientes estatísticos. Esses coeficientes são calibrados por meio de análises regressivas em uma amostra da coleção ou em uma coleção semelhante, para as quais já se tenha julgamento de relevância (COOPER; GEY; DABNEY, 1992). Formalmente ter-se-ia $P(R \mid Q, D)$ a relevância de um documento D para uma consulta Q, assim é proposto na Equação 3.3:

$$\log O(R \mid Q, D) = b_0 + \sum_{i=1}^{S} b_i s_i \wedge P(R \mid Q, D) = \frac{e^{\log O(R|Q,D)}}{1 + e^{\log O(R|Q,D)}}$$

Equação 3.3: *Logistic regression*.

Onde, S é o conjunto de estatísticas gerado a partir da coleção de testes, formado por $\{s_1,....,s_n\}$, b_0 a similaridade do termo com a consulta segundo o modelo probabilístico, e b_i a regressão até *i-1* de *logO(R | Q,D),* na coleção de treinamento.

Após definir os documentos relevantes utilizando a Equação 3.3, é executada a pseudorrealimentação de relevantes para não envolver avaliação por pessoas. Entretanto, uma vez que a proposta não possui variável específica para esse fim, ela é proposta da seguinte maneira: definem-se os 10 documentos inicialmente mais relevantes para cada consulta, definem-se os termos mais relevantes nestes, destes seleciona-se os 10 termos mais relevantes que não constem da consulta através do cálculo do IDF de cada termo. Esses termos são adicionados à consulta, executa-se novamente o algoritmo, retornando à lista definitiva de relevantes (LARSON, 2008).

Porém, esse trabalho utiliza-se unicamente de um TA para o mapeamento de RI-ML, o que nesse caso levou a um percentual de 59,88% em relação à versão monolíngue no melhor caso, obtendo MAP de 35,31% na versão monolíngue e MAP de 21,11% na versão bilíngue francês para inglês.

3.2.5 *Depok*

Utiliza-se de um ou mais dicionários intermediários, permitindo a tradução de uma consulta no idioma A para um idioma B. O trabalho é especialmente interessante, pois compara os resultados de tradução direta e indireta[6], bem como a realimentação de relevantes (HAYURANI; SARI; ADRIANI, 2008).

A Tabela 3.1 compara a técnica de tradução direta da consulta (TA) com a tradução seguida da realimentação de relevantes. A realimentação de relevantes só se mostrou eficiente quando há um grande número de termos a partir de 8 termos na consulta.

Tabela 3.1: Comparativo com expansão de consultas segundo *Depok* (HAYURANI, SARI; ADRIANI, 2008).

Tarefa	Mono	TA	TA + expansão
Título	0,3835	0,3418	0,3375
Título e descrição	0,4056	0,3237	0,3878

Também foi mensurada no trabalho a utilização da tradução transitiva com 1 ou 2 idiomas intermediários. Nos testes foram utilizados os idiomas alemão e francês. Os resultados são demonstrados na Tabela 3.2.

Os resultados do trabalho dão um balizamento da contribuição de expansão de consultas, bem como das perdas causadas pela inserção de um idioma intermediário na tradução. Esse trabalho comprovou que o menor índice de perdas ocorre quando da associação de tradução automática com expansão de consultas. O trabalho também conclui que a realimentação de relevantes é mais eficiente quanto maior o número de pseudorrelevantes selecionados. O trabalho, porém, não efetuou testes com mais de 200 pseudorrelevantes.

[6] Tradução indireta consiste na tradução de um idioma A para um idioma B utilizando um idioma C como intermediário, esta é comum quando não há tradutores entre A e B.

Tabela 3.2: Avaliação de tradução transitiva (HAYURANI; SARI; ADRIANI, 2008).

Tarefa	Monolíngue	Tradução transitiva	Percentual do Monolíngue
Título e descrição	0,4056	0,2831 Unindo as traduções dos dois idiomas	69,80%
Título e descrição		0,3437 Interseccionando os resultados realimentando os 5 primeiros relevantes	84,74%
Título e descrição		0,3297 Interseccionando os resultados realimentando os 10 primeiros relevantes	81,31%
Título e descrição		0,3342 Utilizando alemão como língua intermediária	88,50%
Título e descrição		0,3342 Utilizando alemão como língua intermediária realimentando 10 relevantes	85,31%

3.2.6 *Pattern Matched Translation*

Esse trabalho trata de um assunto ainda pouco explorado em RI-ML, as palavras não processáveis pelas propostas de RI-ML. Em todas as abordagens podem ocorrer termos que não constam de dicionários, de tesauros ou de corpora paralelos. Técnicas para solucionar esse problema foram propostas baseadas em estatística (CHENG et al., 2004; ZHANG; VINES; ZOBEL, 2005) ou em abordagens linguísticas (LIU; JIN; CHAI, 2005). A proposta baseia-se no lançamento do termo não encontrado em um motor de busca para a *Web*. Deste são extraídos os textos dos 500 primeiros resultados. Nesses termos são comparados o padrão de ocorrência do termo buscado com padrões de ocorrências de outros termos, que também não constem da lista do dicionário, ou corpus paralelo (ZHOU et al., 2007). Como forma de melhorar os resultados, o trabalho também propõe a busca por padrões dentro dessas páginas, como termos entre parênteses próximos à palavra buscada que poderiam caracterizar uma tradução.

Nos experimentos, a utilização dessa técnica de detecção de termos fora de dicionário melhorou os resultados em 21% se comparados à versão bilíngue sem a técnica. Entretanto, o resultado ficou em 77% do resultado monolíngue (ZHOU; TRURAN; BRAILSFORD, 2007).

3.2.7 Comparativo dos trabalhos de RI-ML baseados em dados bibliográficos

Os sete trabalhos analisados baseiam-se no mapeamento da consulta de um idioma origem para o idioma dos documentos, assim reduzindo a RI-ML a um problema de RI. Entretanto, as técnicas utilizadas para RI variam de caso a caso. Assim, um índice que passa a ser

importante é o desempenho em relação à *baseline* monolíngue, ou seja, quanto uma abordagem degrada o resultado da consulta, se comparado com sua versão monolíngue. A Tabela 3.3 sumariza a porcentagem da versão bilíngue em relação à monolíngue. O trabalho Oromo-English não apresentou resultados monolíngues, entretanto foi incluído o bom resultado obtido.

Tabela 3.3: Monolíngue X multilíngue.

Técnica	Monolíngue	Multilíngue	Percentual do Monolíngue
Xtrieval	35,72%	34,16%	**95,90%**
XRCE	34,66%	28,25%	**81,49%**
WikiTranslate	34,07%	22,78%	**66,87%**
Logitic Regression	35,31%	21,11%	**59,79%**
Depok	40,56%	38,78%	**95,70%**
Pattern Matched Translation	26,96%	22,55%	**77,00%**

A Tabela 3.4 faz um apanhado das técnicas de RI-ML aplicadas a cada um dos trabalhos submetidos ao CLEF 2008. Já a Tabela 3.5 mostra as técnicas de RI aplicadas em cada trabalho analisado.

Tabela 3.4: Técnicas de RI-ML utilizadas pelas propostas CLEF 2008.

Técnica	Tradução Automática	Tesauro	Dicionário	Corpus paralelo
Xtrieval	✓	✗	✗	✗
XRCE	✗	✓	✓	✓
WikiTranslate	✗	✓	✓	✓
Logitic Regression	✓	✗	✗	✗
Depok	✗	✗	✓	✓
Pattern Matched Translation	✗	✗	✓	✗

Tabela 3.5: Técnicas de RI utilizadas pelas propostas CLEF 2008.

Técnica	Motor	Remoção de *Stopword*	*Stemming*	Expansão de consultas	Realimentação de relevantes
Xtrieval	Lucene	✓	✗	✗	✓
XRCE	Lemur	✓	✗	✓	✓
WikiTranslate	Lucene	✓	✓	✓	✓
Logistic Regression	Próprio	✓	✓	✗	✓
Depok	Lemur	✓	✗	✗	✓
Pattern Matched Translation	Lemur	✓	✓	✗	✗

As propostas *Xtrieval* e *Logistic Regression* possuem foco em RI monolíngue, participando da trilha de RI-ML do CLEF, com foco principal em demonstrar a aplicabilidade da técnica sobre diferentes idiomas, uma vez que o mapeamento de conceitos é feito utilizando recursos amplamente difundidos. Já as técnicas XRCE e *WikiTranslate* possuem foco no mapeamento dos conceitos entre um par de idiomas. Elas se baseiam em corpus paralelo, sendo que delas apenas a *WikiTranslate* utiliza um corpus comparável.

No capítulo seguinte é apresentada a técnica proposta neste trabalho, que juntamente com as demais apresentadas neste capítulo, participou da campanha de avaliação de 2008 do CLEF.

4 APLICAÇÃO DE REGRAS DE ASSOCIAÇÃO A RI-ML

Entre as diversas abordagens para RI-ML, este trabalho optou pela baseada em corpora paralelos. As propostas baseadas em corpora visam sanar os problemas encontrados nas outras abordagens: a inexistência de tesauros abrangentes para muitos pares de línguas; as limitações de dicionários, como a ausência de formas variantes das palavras e da seleção de termos dentre diversas possibilidades disponibilizadas para traduções de um termo; e as traduções errôneas comuns na tradução automática em função da escolha de apenas uma tradução para cada termo.

A partir do mapeamento da relação entre os termos, é possível construir um "tesauro" automaticamente. Essa relação pode ser buscada em dois tipos de corpus: os comparáveis, que tratam do mesmo assunto, mas que não são traduções diretas um do outro, por exemplo, artigos da Wikipédia; e os paralelos, que são compostos por versões de um documento em um idioma origem e em um idioma destino, como, por exemplo, traduções juramentadas. A opção pela utilização de corpora paralelos alinhados visa simplificar o problema em relação a corpora comparáveis.

A geração desse corpus pode ser feita por meio de alinhamento de documentos traduzidos manualmente ou artificialmente com a tradução de documentos que virão a formar um corpus paralelo artificial. Ressalta-se que a qualidade do resultado está diretamente relacionada à qualidade da tradução e ao tamanho do corpus (MCNAMEE; MAYFIELD, 2003). Entre os trabalhos que utilizam corpora paralelos, podemos citar *XRCE* (CLINCHANT; RENDERS, 2009) e *WikiTranslate* (NGUYEN et al., 2009). Destes, o primeiro baseia-se em corpora paralelos alinhados, enquanto que o *WikiTranslate*, em corpora comparável. No entanto, para poder trabalhar com corpora comparáveis, o *WikiTranslate* utiliza-se de *meta-tags* que indicam a tradução de vários termos, o que na maioria das vezes não estaria disponível, sendo comum apenas na Wikipédia e sistemas similares. Já a técnica XRCE utiliza métodos estatísticos complexos para determinação de equivalência semântica entre dois termos, o que possui elevado custo computacional.

Como uma alternativa que possui um tempo de execução bastante baixo, este trabalho propõe a utilização de mineração de dados (MD), na qual idealmente associará um termo as suas possíveis traduções de forma ponderada, ou seja, diria que o termo "A" no idioma origem, é traduzido 70% das vezes como "B" e 30% das vezes como "B2" no idioma destino.

A definição de MD, segundo Fayyad e Uthurusamy (1996), é: "processo de identificação de padrões válidos, inovadores potencialmente úteis e principalmente compreensíveis em conjuntos de dados." A MD é o processo por meio do qual se busca por padrões consistentes e normalmente ocultos (associações, séries temporais ou relacionamentos sistemáticos) em grandes coleções de dados, identificando, assim, subconjuntos de dados com características próprias.

O processo de MD é normalmente dividido em etapas. Vários autores definem números distintos de etapas para esse processo. Fayyad e Uthurusamy (1996) dividem o processo em

nove etapas; Weiss o faz em quatro. Entretanto, a divisão feita por Rezende (2008) é a que se mostra mais apropriada à proposta. Nela, o processo é dividido em três etapas. Essa divisão se mostra apropriada, pois permite uma relação direta com as etapas de RI e de forma mais intrínseca ainda com as etapas de RI-ML. As etapas propostas por Rezende são: pré-processamento, extração de padrões e pós-processamento. Adicionalmente a essas etapas, é necessária a identificação do problema. Também é importante definir previamente o objetivo da MD. A Figura 4.1 ilustra esse processo. A seguir detalharemos as etapas da MD.

Figura 4.1: Etapas da mineração de dados.

4.1 Etapas da Mineração de Dados

4.1.1 Identificação do problema

Consiste na escolha do domínio da aplicação e na definição do objetivo a ser alcançado. Para isso, são definidos os dados necessários para a MD. Nessa etapa também podem ser definidos valores de parâmetros de mineração inicial, os quais poderão ser refinados na etapa de extração (FAYYAD; UTHURUSAMY, 1996). Nesse momento é fundamental uma compreensão sobre a área de atuação do algoritmo, pois isso é fundamental para a boa execução das demais partes do algoritmo, bem como sobre as estruturas dos dados e como elas podem ser processadas.

Para aplicação de MD em RI-ML, o corpus paralelo passa a ser o objeto da mineração, tendo como objetivo a obtenção de equivalentes semânticos de qualquer termo ‘A’ de um idioma origem num idioma destino. Esse corpus paralelo possui frases alinhadas na língua origem com a sua respectiva tradução na língua destino, o que para o algoritmo de mineração será entendido como uma tupla. Cada uma das palavras dessas frases alinhadas é considerada um atributo dessa tupla, que por esse motivo possuirá quantidade variável de atributos.

4.1.2 Pré-processamento

Normalmente, os dados objeto da MD diferem do formato ideal para estes. Além disso, muitas vezes a quantidade de dados é demasiadamente grande para o processamento. Na etapa de pré-processamento são executados métodos para a redução de volume e eliminação de ruídos na seleção de dados. Esse pré-processamento necessita estar em conformidade com os objetivos da MD. Outro fator importante é que esse pré-processamento não apresente um conjunto de dados que não distorça os resultados quando da redução ou limpeza do mesmo.

A seguir detalham-se as etapas do pré-processamento e como estas foram ou não aplicadas à proposta:

- **Extração e integração**

Consiste na obtenção do corpus objeto da mineração e na transformação dele para o formato utilizável pelo algoritmo de mineração.

No trabalho proposto utilizam-se corpora paralelos para essa MD. Como foi identificada em testes, a utilização de unidades menores reduz bastante o ruído. Assim, o alinhamento do corpus paralelo foi feito por sentença.

- **Limpeza**

Apesar de após a limpeza os dados estarem prontos para a MD, um processo de limpeza pode trazer melhores resultados por meio da retirada de dados que possam interferir no processo, como conjuntos contendo dados discrepantes, dados incompletos ou ruidosos. No caso de corpus paralelo, efetuou-se a retirada de pontuações e caracteres acentuados, como é comum na etapa de pré-processamento, aplicada também a RI-ML e remoção de *stop words*, as quais levariam a regras errôneas, pois seriam associadas às traduções de muitos termos.

- **Redução dos dados**

Devido a limitações de tempo ou de capacidade de processamento, por vezes é necessário limitar a quantidade de dados a ser processada. Segundo Weiss e Indurkhya (1998), existem três abordagens para essa redução de volume: redução do número de registros, redução do número de atributos e redução do número de valores de um atributo. Todas essas possibilidades são utilizadas neste trabalho.

A redução do número de registros consiste na retirada de uma amostra do conjunto mantendo as suas características por meio de uma seleção aleatória. Em todos os testes em que foram feitas reduções do número de exemplos, foram extraídos de 20% a 25% dos registros para sobre eles aplicar o algoritmo de mineração.

Redução de número de atributos consiste na seleção de um subconjunto de atributos existentes procurando não impactar ou o fazê-lo de forma mínima na qualidade do resultado. Em todos os nossos testes foram removidas *stop words*, as quais podem ser consideradas atributos de baixa relevância.

- **Redução do número de valores de um atributo**

Pode ser feita por meio da discretização ou suavização de valores. A primeira consiste na transformação de um valor numérico em informações de intervalo, os quais são agrupados, tais como de 1 a 3 baixo, de 4 a 6 médio, acima disso, alto. A segunda consiste também na substituição de um grupo de valores por um novo valor, que poderia ser uma média deles. No caso da aplicação de MD, a RI-ML pode ser considerada como uma suavização do processo de *stemming*, pelo qual passam todos os termos do corpus paralelo.

Como foi exemplificado, há diversas semelhanças entre o pré-processamento de MD e de RI-ML. Apesar de o primeiro normalmente ser aplicado comumente sobre valores numéricos ou binários, suas etapas também se mostraram aplicáveis a textos, foco da RI-ML.

4.1.3 Extração de padrões

A extração de padrões consiste na escolha e na aplicação da tarefa de MD e do algoritmo de MD propriamente dito. No entanto, é necessário definir qual(is) algoritmo(s) será(ão) utilizado(s) e quais os parâmetros passados. Os algoritmos de MD podem ser classificados em dois grandes grupos: os preditivos e os descritivos.

A MD preditiva parte de dados anteriores de treinamento previamente obtidos para exemplificar grupos. Os problemas de predição são classificação e regressão. O primeiro atribui ou não a um grupo uma dada tupla, por exemplo, o cliente "X" é classificado como 4 estrelas por um dado banco. Os problemas de regressão trabalham com intervalos numéricos, por exemplo, a predição do lucro médio de um dado fundo de investimento (WEISS; INDURKHYA, 1998).

A MD descritiva busca por comportamentos sistemáticos dentro do conjunto de dados. As tarefas principais são *clustering*, RAs e sumarização. Os algoritmos de *clustering* fazem agrupamentos automáticos de dados segundo seu grau de semelhança. O critério de semelhança faz parte da definição do problema e depende do algoritmo (WIKIPÉDIA, 2010). Já a sumarização é o processo de produção de uma versão reduzida de um texto, geralmente pela seleção e estruturação de seu conteúdo informativo mais relevante (SPARCK-JONES, 1993). A seção 4.2 trata das RAs, pois este foi o grupo de algoritmos que se mostrou útil para a RI-ML.

Uma vez definido o tipo de mineração de dados a ser executada, deve-se escolher um algoritmo apropriado. Esse processo será mais bem detalhado na seção 4.2 referente a RAs.

4.1.4 Pós-processamento

No pós-processamento os dados serão avaliados e visualizados. Também é o momento da eliminação de padrões, que, apesar de atenderem aos requisitos da MD, não atendem aos interesses da resposta requerida, nesse caso, a tradução de consultas. Segundo Bruha e Famili (2000), os procedimentos de pós-processamento podem ser agrupados nas seguintes categorias:

- Filtragem do conhecimento: definem padrões de poda em árvores de decisão, ou restrições em outros algoritmos.
- Interpretação: torna o conhecimento compreensível ao ser humano ou ao sistema para o qual será utilizado em um fluxo de execução.

- Avaliação: define a eficiência e precisão dos padrões gerados, e integração quando da utilização de vários métodos de RI-ML, assim definindo como o resultado de todas elas será composto.

Essas etapas de pós-processamento posteriormente serão aplicadas ao resultado do algoritmo escolhido. Entretanto, aqui novamente é possível traçar um paralelo com RI-ML que tem como atividade muito importante a avaliação de resultados. A atividade de filtragem, que no caso de RI-ML, consiste na atribuição de equivalência semântica entre termos em idiomas, uma das principais contribuições deste trabalho, é detalhada na seção 4.7.

4.2 Regras de associação

Dentro da MD, a técnica baseada em RAs pode ser classificada como pertencente ao modelo descritivo (AGRAWAL; SRIKANT, 1994). As RAs têm como objetivo encontrar um conjunto de itens frequentes em registros ou transações e identificar a influência desses conjuntos na presença de um outro conjunto.

Formalmente, uma RA pode ser definida como: dado um subgrupo de características presentes em um subconjunto *C* de dados de um conjunto *U*, estes levam à existência de um outro subgrupo de dados com características frequentes em *C*, identificando tendências de correlações existentes sobre uma coleção de dados. Exemplos no cotidiano da apresentação dessas técnicas estão disponíveis em *sites* como Amazon[7], Submarino[8] etc., nos quais há indicações de: "Quem compra o produto *P* também compra o produto *X*". Nesse exemplo, *U* seria o banco de dados da Amazon; *C*, por sua vez, seria o conjunto de pessoas que compraram *P*; assim, X seriam tendências identificadas dentro do subconjunto *C*.

Uma implementação clássica de mineração de dados por RAs é atribuída a Agrawal (AGRAWAL; SRIKANT, 1994). Como o algoritmo foi originalmente concebido para mineração de regras sobre bancos de dados, ele é projetado para trabalhar sobre relações. A partir dessas relações o algoritmo gera agrupamentos de itens frequentes. Esses agrupamentos são conjuntos de itens ordenados lexicograficamente (AGRAWAL; SRIKANT, 1994). Visando evitar a explosão combinatória, é definido um tamanho máximo para esses agrupamentos, bem como um suporte mínimo, a seguir explicado.

Para iniciar o processo de mineração, cada item $X \subseteq A$ (sendo *A* o conjunto de todos os itens da coleção) é considerado um possível nó nascedouro de um grafo de saída do algoritmo, à exceção do conjunto vazio. Porém, essa opção, mesmo para coleções pequenas, geraria grafos com grande quantidade de nodos. Para reduzir esse problema, o algoritmo minerador de RAs define a linha de limiar, que consiste em definir uma quantidade mínima de ocorrências de um dado agrupamento de itens. Agrawal e Srikant (1994) definem que todo subconjunto de um agrupamento de itens frequentes deve também ser frequente. Assim, o algoritmo pode ciclicamente calcular o suporte e remover agrupamentos de itens que tenham ao menos um subgrupo não presente no grupo dos itens frequentes. Assim, por exemplo, em uma busca sobre uma coleção de 10.000 tuplas com suporte de 5%, ao final da primeira iteração do algoritmo, ou seja, quando só existirem agrupamentos de tamanho 1, serão eliminados todos os itens que possuírem ocorrência em menos de 500 itens - chamaremos essa coleção de K_1. Após esse passo, o algoritmo é repetido procurando sequências de tamanho 2 em que todas as sequências de tamanho 1 estejam na coleção K_1. Aplica-se

[7] Disponível em: <http://www.amazon.com>.
[8] Disponível em: <http://www.submarino.com.br/>.

novamente o limiar de suporte, desta vez buscando sequências de tamanho 2 com o suporte desejado. E esse processo inicia novas iterações até um limite *N,* definido como parâmetro de entrada do algoritmo. Apesar de existirem diversas implementações de algoritmos de mineração de RAs, todas elas devem gerar o mesmo resultado.

As principais métricas em sistemas de RAs para mineração de dados são o suporte e a confiança. Sendo *N* o número total de tuplas da relação, *P* um conjunto ordenado que represente o precedente da regra, e *C* o consequente, define-se:

Suporte como a frequência relativa em que uma regra $P \Rightarrow \mathrm{C}$ ocorre na coleção de tamanho N, e onde $n(P \cup \mathrm{C})$ é o total de transações nas quais *P* e *C* ocorrem juntos, conforme demonstrado na Equação 4.1.

$$\sup(\mathrm{P} \Rightarrow \mathrm{C}) = \sup(\mathrm{P} \cup \mathrm{C}) = \frac{\mathrm{n}(\mathrm{P} \cup \mathrm{C})}{N}$$

Equação 4.1: Suporte em regras de associação.

Confiança como a possibilidade de *C* ocorrer dada a ocorrência de *P,* conforme demonstrado na Equação 4.2.

$$\mathrm{conf}(\mathrm{P} \Rightarrow \mathrm{C}) = \frac{\sup(\mathrm{P} \Rightarrow \mathrm{C})}{\sup(\mathrm{P})} = \frac{n(\mathrm{P} \Rightarrow \mathrm{C})}{n(\mathrm{P})}$$

Equação 4.2: Confiança em regras de associação.

Com essas medidas é possível definir os passos do algoritmo de mineração, como: inicialmente definir todos os conjuntos de itens nos quais $\sup(P \Rightarrow \mathrm{C})$ é maior que o limiar definido na entrada, e tamanho é maior que 1 e menor ou igual a um K, chamando esse conjunto de *S*. Para cada subconjunto $I \subseteq S$, para todo subconjunto $\tilde{v} \subseteq I$, gerar uma regra na forma $\tilde{v} \Rightarrow (I - \tilde{v})$, então se $\sup(I)/\sup(\tilde{v})$ é maior do que o suporte mínimo, este é adicionado ao conjunto de saída.

Na Figura 4.2 é demonstrada a execução do algoritmo, considerando os termos *a, b, c, d, e* como termos de uma coleção. A partir destes são feitas combinações de todos os termos, inicialmente dois a dois e posteriormente três a três, e assim sucessivamente. Vale ressaltar que segundo o algoritmo, alguns resultados abaixo da linha de suporte mínimo não seriam calculados devido ao descarte no caso em que um dos subconjuntos não possua o suporte mínimo necessário. Por fim, seriam retornados apenas os resultados acima da linha de suporte.

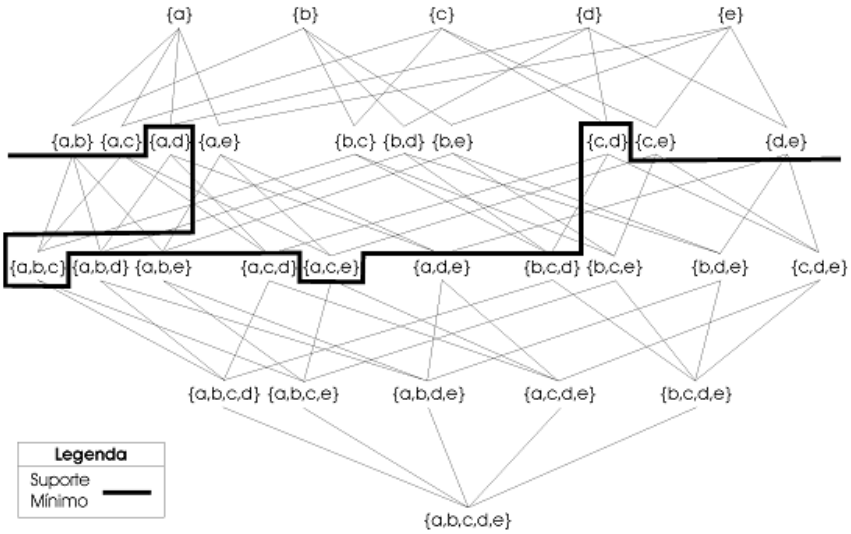

Figura 4.2: Corte por nível de suporte.

Além do suporte e confiança mínima, também se pode adicionalmente definir suporte e confiança máximos, por regras análogas.

Como exemplo de RAs podemos definir os conjuntos de itens: {A,B,C,D}, {A,B,D}, {B,C},{A,C,D},{B,C,D} e {A,B,C,D}, em que cada letra pode corresponder a um termo de uma frase. O primeiro passo do algoritmo de RAs é contar as frequências de cada item separadamente. Isso dará o suporte para as regras de tamanho 1. Assim, teríamos:

Tabela 4.1: Conjunto inicial de itens.

Item	Suporte
A	4
B	6
C	4
D	5

A partir deste nível, pode-se definir um valor mínimo de suporte para o corte por nível de suporte, assim define-se para exemplificar o valor 4. Dessa forma, apenas os subconjuntos com quatro ocorrências apareceriam, e teremos os seguintes valores (os valores tachados foram podados pelo algoritmo de RAs):

Tabela 4.2: Primeira filtragem.

Item	Suporte
{A,B}	4
~~{A,C}~~	~~2~~
{A,D}	4
{B,C}	4
{B,D}	5
~~{C,D}~~	~~3~~

Todos os conjuntos como {A,C} e {C,D} que não atingiram o suporte mínimo para regras de tamanho dois não serão testados com regras de tamanho três. Assim, por testes dos subconjuntos restantes com todos os possíveis itens, teríamos os seguintes conjuntos de tamanho três:

Tabela 4.3: Segunda filtragem.

Item	Suporte
{A,B,D}	4
~~{B,C,D}~~	~~3~~

Em sistemas com grande volume de dados, como textos em corpora paralelos, a poda reduz drasticamente a quantidade de processamento necessária, pois os níveis de coocorrências de palavras são bastante baixos. Palavras bastante comuns, entretanto, que não se enquadram em *stopwords*, ocorrem em menos de 2% das linhas.

4.3 Algoritmo Apriori

Criado por Agrawal e Skrikant (1994), ele segue todas as definições acima citadas para algoritmos de mineração de RAs, tendo-se diferenciado pela performance aliada à simplicidade de sua implementação original.

Dada uma base de dados D composta por itens $I = \{i_1, ..., i_m\}$ e um conjunto de transações $T = \{t_1, ..., t_n\} / (t_i \in T \wedge t_i \subseteq I)$, o algoritmo segundo Agrawal e Skrikant (1994) é conforme a Figura 4.3:

1) $L_1 := \{1\text{-itemsets frequentes}\}$
2) **for** $(k := 2; L_{k-1}; k := k+1)$ **do**
3) $C_k := \text{apriori-gen}(L_{k-1})$;
4) **for all** (transações $t \in T$) **do**
5) $C_t := \text{subset}(C_k, t)$;
6) **for all** candidatos $c \in C_t$ **do**
7) c.count := c.count + 1;
8) **end for;**
9) **end for;**
10) $L_k := \{c \in C_k \mid c.count \geq \textbf{sup - min}\}$;
11) **end for**
12) Retorna := $\cup_k L_k$

Figura 4.3: Algoritmo Apriori.

Segundo Gillmeister e Cazella (2007), o funcionamento do algoritmo Apriori está resumidamente demonstrado na Figura 4.3. No início, L_1, que é o conjunto de grupos com somente um elemento, é gerado. Na sequência, tem-se um laço com *k* passos. Neste serão desenvolvidas basicamente duas tarefas. A primeira é a geração do grupo de itens candidatos C_k , através dos itens gerados no passo anterior (conjunto de L_{k-1}) e utilizando-se da função Apriori-gem para isso. A segunda tarefa executada no laço *k*, consiste num outro laço para contagem do suporte dos itens (*c*) do grupo candidato C_k , em que cada transação da base de dados é analisada. Neste momento, o Apriori utiliza uma função estruturada na forma de uma *hash-tree*, na qual cada nodo folha contém uma lista de itens ou o endereçamento para uma tabela *hash*. Assim, é possível encontrar todos os candidatos contidos na transação *t* de forma ágil. Cada candidato *c* terá ao final o seu suporte computado, e no próximo passo *k*, os itens que não obtiveram o suporte mínimo estabelecido são excluídos.

4.4 RI-ML utilizando RAs

As semelhanças existentes entre RAs e RI-ML permitem traçar um paralelo entre ambas as técnicas. Assim, é possível comparar o pré-processamento de RI-ML com o de RAs. A etapa de mineração é análoga ao mapeamento de conceitos entre pares de idiomas em RI-ML, e o pós-processamento de RAs pode ser comparado às etapas de realimentação de relevantes e expansão de consultas.

A Figura 4.4 mostra as fases da proposta apresentada nesta dissertação quanto ao uso de RAs para RI-ML. O processo é dividido em 5 fases que serão descritas nas próximas subseções.

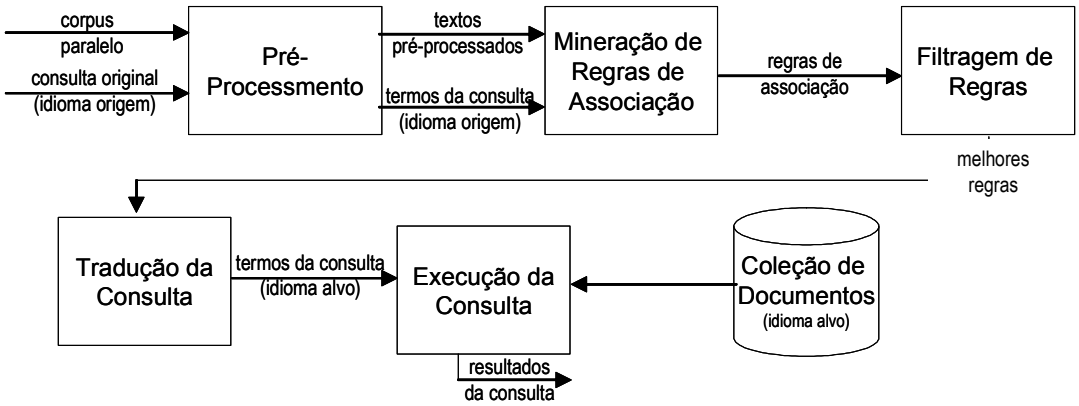

Figura 4.4: Etapas de RI-ML com regras de associação.

4.4.1 Pré-processamento em RI-ML utilizando RAs

Na etapa de pré-processamento pode-se utilizar o pré-processamento de RI-ML. É executada a limpeza dos dados por meio da remoção de caracteres de controle como marcadores de campos e marcadores gramaticais tais como pontuação e acentuação. Por exemplo, além do mapeamento de letras maiúsculas para minúsculas, também se exclui quaisquer sequências de caracteres que não possam caracterizar palavras, ou seja, que possuam caracteres que após remoção de marcadores gramaticais difiram de [a-z], dessa forma diminuindo ruído e também o tamanho da coleção.

Nessa etapa ainda são executadas a remoção de *stopwords* e *stemming*. Entretanto, a partir de experimentos em menor escala, verificou-se a necessidade de diminuir o tamanho das sentenças. Assim, as sentenças normalmente alinhadas por parágrafos passaram a ser por linhas, dessa forma, o número de erros nas etapas seguintes é diminuído. Outra adição importante nesta etapa é a de *meta-tags* no corpus paralelo que permitam definir em que idioma está um termo. Essa adição objetiva a eliminação na etapa de pós-processamento de regras que envolvam apenas um idioma, o que caracteriza termos compostos ou comumente coocorrentes, mas não possíveis traduções para um termo. Ressalta-se que a adição de *meta-tags* e o alinhamento das sentenças acontecem apenas no corpus paralelo que será utilizado para a extração das RAs. As demais etapas devem ser aplicadas a todas as etapas do processo.

4.4.2 Mineração de regras de associação

Mapeamento semântico é o principal objetivo da RI-ML. Ele consiste no mapeamento de um termo num dado contexto de um idioma origem para um ou mais termos no idioma destino. Para tanto, este trabalho propõe a utilização de um corpus paralelo como base para geração de um dicionário, que possui como principal vantagem, em relação aos dicionários prontos, possuir a seleção ponderada de diversos termos para uma mesma palavra no idioma origem, evitando, dessa forma a seleção de apenas um termo.

A fim de diminuir o tamanho do corpus de busca e evitar a explosão combinatória resultado das combinações das dezenas de milhares de termos existentes em uma coleção, são extraídas apenas as frases que possuem cada um dos termos da busca. Por exemplo, para uma consulta “Anistia internacional América Latina”, criam-se quatro subcorpora (um para cada termo da consulta), cada um dos quais possuindo apenas as sentenças nas quais a respectiva palavra ocorre. Dessa forma, é possível ao algoritmo de mineração procurar apenas por regras que ocorram em todas as sentenças, ou seja, onde o suporte é 100%, assim retornando a

relação do termo com todos os demais termos da coleção, dentro do limite de confiança estipulado. Isso, além de simplificar o pós-processamento, adiciona um grande ganho de performance ao sistema apesar do custo computacional do fracionamento do corpus, que é diretamente proporcional ao número de termos distintos da consulta. Entretanto, o custo computacional da proposta ainda se mantém baixo devido ao baixo custo de RAs *O(log(N))* (WIJSEN; MEERSMAN, 1998), comparando-se, por exemplo, com a complexidade da técnica de Indexação Semântica Latente que utiliza SVD, cuja complexidade é $O(min(NM^2,NM^2))$, onde N é o total de termos distintos da coleção e M o total de documentos.

4.4.3 Filtragem das regras de associação

Após a criação dos subcorpora ocorre a filtragem de regras, que consiste na etapa mais complexa do trabalho. A filtragem é feita com heurísticas determinadas com base na observação das regras geradas. Para os testes realizados elas mostram uma melhoria no desempenho. A filtragem é composta pelas seguintes etapas:

- Define-se de um valor mínimo de confiança. Regras com confiança inferior a esse valor não serão retornadas.
- Eliminam-se as regras com ambos os termos no mesmo idioma, pois o objetivo é o mapeamento de termos entre idiomas.
- Seleciona-se o elemento de maior confiança, aqui chamado de **M,** este será indicado como uma das traduções do termo.
- Selecionam-se elementos com confiança até 80% da confiança de **M.** Em caso de termos onde há mais de duas traduções possíveis, mas duas delas são dominantes, este limiar é utilizado. Os resultados de experimentos mostram que essa regra aplica-se em apenas 0,5% dos termos.
- Selecionam-se elementos com confiança igual a (100)-**M**±0,1**,** uma vez definida a tradução mais corriqueira do termo, e supondo-se que o termo sempre será traduzido, o somatório das traduções do termo deve completar 100%.

A Figura 4.5 exemplifica a busca para um termo, nesse caso, o *stem* "civil" em português. Esse *stem* aparece em todas as regras, resultado do subcorpus onde todas as frases o possuem. A representação abaixo possui a seguinte formatação, ***Stem* possível** <- ***stem* origem** (**Suporte**/**Quantidade de ocorrências**, **confiança**), uma vez que o suporte é definido como 100%. Sobre as regras geradas, aplicou-se heurísticas de filtragem.

```
~~E=war <- civil (100.0/960, 26.1)~~          Confiança baixa- eliminado
E=civilian <- civil (100.0/960, 29.6)     Complemento para 100 selecionado
~~guerr <- civil (100.0/960, 25.6)~~          palavras em mesmo idioma eliminado
E=civil <- civil (100.0/960, 70.5)                Melhor confiança selecionado
```

Figura 4.5: Exemplo de filtragem de regras de associação.

Após a execução da filtragem de regras, os prefixos usados como marcadores de idioma são removidos.

4.4.4 Tradução da consulta

Como resultado da etapa de filtragem, temos um conjunto de regras que mapeiam cada termo da consulta para suas traduções mais prováveis. A tradução da consulta é feita substituindo-se cada termo pelos seus equivalentes no outro idioma.

4.4.5 Execução da consulta

Nesta etapa, a consulta traduzida para o idioma dos documentos é submetida ao motor de RI monolíngue. Como saída obtém-se um *ranking* com os documentos recuperados para cada consulta. Com esse *ranking* e os julgamentos de relevância fornecidos por campanhas de avaliação como a CLEF, é possível avaliar a qualidade do resultado do sistema.

Se os resultados não forem satisfatórios, pode-se aplicar novamente o ciclo de mapeamento semântico, modificando seus parâmetros de configuração, como a confiança mínima para uma regra ser avaliada.

Cabe ressaltar que a coleção usada como base para a mineração de RAs não precisa ser a mesma utilizada para a recuperação dos documentos. É possível extrair as RAs de um corpus bilíngue e utilizar corpus diferente para teste.

Existem duas estratégias básicas para processar as RAs a fim de criar um dicionário bilíngue: (i) *eager*, que minera regras para todos os termos da coleção e os guarda em um cache; (ii) *lazy*, que minera sob demanda apenas os termos da consulta. Neste trabalho, ambas as abordagens são testadas. No entanto, a opção pela técnica *lazy* é mais produtiva, pois evita a necessidade de novo processamento de todas as regras em caso de aumento do tamanho do corpus paralelo, além de trazer uma melhor qualidade no resultado, uma vez que podem-se definir limites menores de confiança para as regras. Isso diminui o custo computacional total, pois apenas as regras de termos efetivamente buscados são processadas. Por outro lado, essa escolha traz um tempo maior necessário a cada consulta. Dessa forma, a estratégia *eager* foi adotada para as consultas executadas via Internet. Como solução intermediária adotou-se a gravação das regras já pesquisadas para consultas posteriores, desde que neste ínterim não tenham sido adicionados dados ao corpus paralelo.

4.5 Considerações Finais

Este capítulo apresentou a proposta de RI-ML baseada em RAs para corpora paralelos alinhados. Foram abordadas as etapas de RAs e sua semelhança com etapas de RI-ML. Dessa forma, demonstrou-se como aplicar algoritmos de RAs como método para definir mapeamento entre termos em diferentes idiomas.

A principal contribuição deste trabalho é definir um método que permita em um corpus paralelo a seleção de termos em um idioma que representam o mesmo conceito em outro. Entretanto, este trabalho não propõe a seleção de um único termo para essa tradução, mas a utilização dos diversos termos, os quais no corpus paralelo foram utilizados como tradução para o termo da busca.

Este trabalho, portanto, define um método que permite, dado um termo em um idioma A, definir quais todas as traduções deste e suas probabilidades de ocorrência em um determinado corpus paralelo. Dessa forma, o resultado é mais amplo do que o resultado de uma tradução automática, por saber a probabilidade de ocorrência de cada termo, e não ser obrigado a selecionar apenas uma tradução.

No próximo capítulo serão demonstrados experimentos feitos com essa proposta e submetidos a conferências a fim de validar o método, bem como os protótipos desenvolvidos para este trabalho.

5 AVALIAÇÃO EXPERIMENTAL

Neste capítulo são apresentados os experimentos realizados que visam demonstrar que a técnica proposta aproxima o resultado de uma RI monolíngue, bem como o de outros trabalhos da área. A fim de testar a aplicação da proposta em diversos cenários, foram variados:

- o idioma em que as consultas foram executadas;
- o conjunto de consultas, tendo sido avaliadas ao todo mais de 300 consultas;
- os tipos de corpora de busca, desde documentos textuais longos, até metadados;
- o corpus paralelo usado para geração das RAs entre opções como: uma parte do próprio texto, outra coleção do mesmo domínio, corpora de outro domínio traduzidos manualmente.

Dessa forma, demonstra-se que a proposta tem amplitude de aplicação e não é atrelada a um dado corpus, ou características deste, da língua, ou das consultas. Contudo, os experimentos revelam a influência desses fatores sobre a eficiência da proposta.

Este capítulo está dividido da seguinte forma: a seção 5.1 define os recursos utilizados nos experimentos, tais como: métricas de avaliação, plataforma de *hardware* e *software*, corpora, e as formas de alinhar esses corpora. Os experimentos e os resultados são apresentados na seção 5.2. O capítulo é finalizado com uma análise dos resultados na seção 5.3.

5.1 Recursos utilizados

5.1.1 Métricas para avaliação

Nos experimentos serão utilizadas as seguintes métricas para avaliação dos resultados:

- Média das Precisões Médias (MAP): média não interpolada das precisões médias para um conjunto de registros, ver seção 2.1.4 (MANNING; RAGHAVAN; SCHÜTZE, 2008).
- Precisão interpolada: precisão em 11 pontos define a precisão interpolada dos documentos a cada 10% dos documentos recuperados conforme demonstrado na seção 2.1.4 (MANNING; RAGHAVAN; SCHÜTZE, 2008).
- Teste T (*Student's t-test*): verifica se a diferença de desempenho de dois algoritmos é estatisticamente significativa (WIKIPÉDIA, 2009). O teste T avalia se as médias de duas distribuições normais de valores são estatisticamente diferentes. Segundo Hull (1993), o teste T apresenta bons resultados mesmo quando as distribuições não são perfeitamente normais. Foi utilizado o limiar de significância estatística $\alpha = 0{,}05$. Quando o valor do *P* bi-caudal é menor que α, existe diferença significativa entre os

desempenhos dos dois algoritmos analisados. O melhor desempenho é do algoritmo com a maior média de distribuição.

5.1.2 Plataforma de trabalho

Os experimentos foram executados em dois microcomputadores, ambos com a seguinte configuração: Processador *Pentium* 4 2,4GHz, 512 Mb de memória RAM, DDR 266 MHz e discos rígidos de 80 GB 5400 RPM.

Os algoritmos de pré-processamento, MD e RI, foram desenvolvidos em C e apenas adaptados para este trabalho. Os algoritmos de geração de índices invertidos, extração e seleção de regras de associação foram desenvolvidos para o framework.Net, tendo sido parte dele implementada em C# e outra em Object Pascal. Os índices invertidos foram armazenados em banco de dados PostgresSQL 8.3 rodando *Linux Fedora* 4.0 em uma das máquinas. Todos os demais algoritmos rodavam na outra máquina em *Windows XP*.

Com base nessa configuração, o tempo médio para execução de um conjunto de consultas da campanha CLEF é de 12 segundos, desde que os dados já estejam previamente indexados.

5.1.3 Sistema de recuperação de informações

O sistema de RI utilizado nos experimentos foi o *Zettair* (ZETTAIR, 2007). Essa escolha se baseou em testes preliminares nos quais esse sistema demonstrou excelente velocidade para consulta à grande quantidade de dados, além da disponibilidade do código fonte. Esse motor de busca possui diversas métricas de RI para definir similaridade entre consultas e documentos. Utilizou-se a métrica BM25+, a qual propõe modificações em relação à métrica BM25 original, com o objetivo de enfatizar ainda mais termos raros, e melhoras significativas na MAP comparando com outras métricas da ferramenta (GERALDO; ORENGO, 2008).

5.1.4 Coleções de teste

Para validar a técnica proposta, utilizaram-se as coleções das campanhas CLEF no período de 2002 a 2005 e de 2008 (dados de 2006 e 2007 não foram utilizados por não termos acesso a eles). A cada ano são disponibilizadas 50 consultas em diversos idiomas[9]. Cada consulta é composta por título, descrição e narrativa, entretanto apenas título e descrição podem ser utilizados para permitir uma comparação com outros trabalhos. Essas consultas nos diversos idiomas são traduções das mesmas consultas no idioma da coleção de busca, tipicamente o inglês. Também são disponibilizados posteriormente aos participantes os julgamentos dos documentos, enviados por grupos que participaram da campanha como possíveis relevantes.

É importante salientar que essas coleções de testes possuem variações de temas nas consultas que permitem uma boa avaliação da proposta. Para tanto, este trabalho utilizou as três coleções disponibilizadas pelo CLEF no período:

- Notícias do *Jornal Los Angeles Times* do ano de 1994, compostas de 365 edições, totalizando 113.046 notícias (372 Mb de informação). O número de termos distintos é de aproximadamente 113.005, que são resumidos a 90.112 *stems*. Cada documento possui em média 569 termos relevantes, além de 255 estando na lista de *stopwords*. É uma coleção monolíngue escrita na vertente americana do Inglês disponível em formato SGML. Essa coleção foi utilizada pelo CLEF no ano de 2002, e no período de 2004 a 2007 em conjunto com o jornal *Glasgow Herald*.

[9] Exceto em 2003, quando foram disponibilizadas 60 consultas.

- Notícias do Jornal *Glasgow Herald* do ano de 1995, compostas de 312 edições desse jornal disponível em formato SGML, totalizando 56.472 notícias (150 Mb de informação). O número de termos distintos é de aproximadamente 88.874, que são resumidos a 64.495 *stems*. Cada documento possui em média 569 termos relevantes, além de 255 estando na lista de *stopwords*. É uma coleção monolíngue escrita na vertente Britânica do Inglês. Essa coleção foi utilizada pelo CLEF no ano de 2003, e no período de 2004 a 2007 em conjunto com o Jornal *Los Angeles Times*. A Figura 5.1 exemplifica um tópico.
- Metadados da Biblioteca Britânica: essa coleção, ao contrário das demais, é composta por metadados, e não documentos. Trata-se de metadados bibliográficos em formato XML, da Biblioteca Britânica. Sua composição é de 1.000.101 documentos (195 Mb de metainformação). O número de termos distintos é de 689.053. Essa elevada quantia de termos deve-se aos diversos idiomas nos quais os títulos podem estar escritos. Entretanto, cada documento possui apenas 19 termos relevantes em média. Ao contrário das demais, possui diversos campos de informação. Assim, cabe a cada um que a utiliza definir quais campos são ou não relevantes. Tendo sido utilizada pela primeira vez no CLEF de 2008, a Figura 3.2 exemplifica um documento.

Essas coleções foram utilizadas pois possuem avaliação de relevância de resultado para consultas segundo o método de *Pooling* (SPARCK-JONES; RIJSBERGEN, 1975), conforme explicado na seção 3.1.1, além de serem, dentre as disponíveis, as que possuem consultas em maior número de idiomas com tradução oficial da conferencia de avaliação. A isso soma-se a participação dessa proposta no CLEF 2008. Esses corpora possuem variação de vertente idiomática (norte-americana e britânica no caso dos jornais, respectivamente) e o corpus da Biblioteca Britânica. Este último possui documentos multilíngues, com títulos em outros idiomas e com vocabulário muito específico, dado a variabilidade e profundidade dos assuntos em uma biblioteca desse porte comparada a um jornal, que deve ter fácil assimilação pela população mesmo com baixos níveis de escolaridade, trazendo, dessa forma, boa variabilidade de corpora aos experimentos e robustez para ser utilizada em diversos padrões de corpus.

```
<DOC>
<DOCNO>GH950102-000012</DOCNO>
<TEXT>
    1. DAN Marino, playing at home in the Joe Robbie Stadium, outplayed Joe Montana on
       Saturday in a shoot-out between two of American football's greatest
       quarterbacks.
    2. He threw for two touchdowns and the Miami Dolphins capitalised on two late
       turnovers to beat Kansas City 27-17 in the NFL play-offs. AFC East champions
       Miami visit the AFC West champions, San Diego.
    3. In the first Montana-Marino contest since the 1985 finale, Marino completed 22
       of 29 passes for 257 yards. Montana hit 26 of 37 passes for 314 yards and two
       touchdowns.
    4. The Green Bay Packers contained the league's rushing champion, Barry Sanders,
       and held off a late Detroit drive to beat the Lions 16-12 in the NFC wild-card.
    5. And yesterday, Vinny Testaverde passed for 268 yards and one touchdown, leading
       the Cleveland Browns past the New England Patriots 20-13.
    6. The Browns backed Testaverde with a defence that intercepted New England
       quarterback Drew Bledsoe three times.
    7. Cleveland now meet archrivals, Steelers on Saturday in Pittsburgh.
</TEXT>
</DOC>
```

Figura 5.1: Documento original *Glasgow Herald*.

5.1.5 Corpora paralelos

Em virtude da dificuldade em obtenção de corpus paralelo para aplicação do método proposto de RI-ML, necessitou-se a criação de corpora paralelos artificiais. Para as coleções baseadas em notícias dos jornais *Los Angeles Times* e *Glasgow Herald* foram extraídas 20% dos mesmos (por meio da obtenção da terceira, oitava, décima terceira e assim sucessivamente edições anuais). Para o corpus baseado na Biblioteca Britânica foram extraídos 25% dos conjuntos de registros para geração do corpus paralelo. Esse valor maior é necessário em função da maior variedade de vocabulário dessa biblioteca. Em todos os casos, a seleção dos documentos utilizados para geração do corpus paralelo não leva em consideração a relevância do documento, não invalidando, dessa forma, a utilização deste nas consultas. Por fim, para textos baseados nas transcrições de reuniões do parlamento europeu, utilizou-se a totalidade dos textos, pois esses documentos em nenhum momento foram utilizados como corpus objeto da recuperação de informação, nesse caso formando um corpus paralelo traduzido manualmente.

```
<DOC>
<DOCNO>GH950103-000012</DOCNO>
<TEXT>
    1. DAN Marino, playing at home in the Joe Robbie Stadium, outplayed Joe Montana on
       Saturday in a shoot-out between two of American football's greatest
       quarterbacks.
    2. He threw for two touchdowns and the Miami Dolphins capitalised on two late
       turnovers to beat Kansas City 27-17 in the NFL play-offs.
    3. AFC East champions Miami visit the AFC West champions, San Diego.
    4. In the first Montana-Marino contest since the 1985 finale, Marino completed 22
       of 29 passes for 257 yards.
    5. Montana hit 26 of 37 passes for 314 yards and two touchdowns.
    6. The Green Bay Packers contained the league's rushing champion, Barry Sanders,
       and held off a late Detroit drive to beat the Lions 16-12 in the NFC wild-card.
    7. And yesterday, Vinny Testaverde passed for 268 yards and one touchdown, leading
       the Cleveland Browns past the New England Patriots 20-13.
    8. The Browns backed Testaverde with a defence that intercepted New England
       quarterback Drew Bledsoe three times.
    9. Cleveland now meet archrivals, Steelers on Saturday in Pittsburgh.
</TEXT>
</DOC>
```

Figura 5.2: Documento com separação por frases.

Após a seleção dos documentos que serão traduzidos, estes têm as frases de seus parágrafos separadas. A Figura 5.2 exemplifica um documento em seu formato original. Essas frases passam a ser traduzidas pelo tradutor *on-line* da *Google* (GOOGLE, 2007), utilizando computação nas Nuvens. A Figura 5.3 exemplifica a tradução para português do documento contido na Figura 5.2, gerando, dessa maneira, dois documentos com o mesmo número de linhas e passíveis de alinhamento. O idioma para o qual os documentos originalmente em inglês são traduzidos varia em função do experimento para testar a eficiência da técnica com diferentes idiomas.

```
<DOC>
<DOCNO>GH950103-000012</DOCNO>
<TEXT>
    1.  Dan Marino, jogando em casa, no Estádio Joe Robbie, Joe Montana outplayed no
        sábado em um shoot-out "entre dois dos maiores zagueiros do futebol americano.
    2.  Ele jogou para dois touchdowns e do Miami Dolphins capitalizada em dois
        turnovers tarde para bater Kansas City 27-17 na NFL play-offs.
    3.  AFC East campeões Miami visitar o AFC West campeões, San Diego.
    4.  No primeiro Montana-Marino concurso desde o final 1985, Marinho completou 22 de
        29 passes para 257 jardas.
    5.  Montana acertar 26 de 37 passes para 314 jardas e dois touchdowns.
    6.  O Green Bay Packers continha o campeão da Liga da pressa, Barry Sanders, e
        realizada fora de uma tarde de carro até bater o Detroit Lions 16.-12. no CNF
        wild-card.
    7.  E ontem, Vinny Testaverde passou para 268 jardas e um touchdown, levando o
        Cleveland Browns passado, o New England Patriots 20-13.
    8.  O Browns Testaverde apoiado com uma defesa que interceptadas Nova Inglaterra
        quarterback Drew Bledsoe três vezes.
    9.  Cleveland agora reunir archrivals, no sábado, em Pittsburgh Steelers.
</TEXT>
</DOC>
```

Figura 5.3: Documento traduzido com separação por frases.

Antes de efetuar o alinhamento é necessário um pré-processamento que consiste na remoção de pontuação, passagem de todo o documento para caracteres minúsculos, remoção de *stopwords* e sufixos dos termos, e na adição de um marcador que permita diferenciar um termo em um idioma do outro. Em todos os experimentos foi utilizada a *string* "E=" para identificar documentos em inglês, já que em todos os trabalhos esse idioma é envolvido em função das coleções de teste estarem todas disponíveis nele. Por fim, os textos no idioma origem e destino são alinhados, como é mostrado na Figura 5.4.

Paralelamente a essa tarefa, a mesma etapa de pré-processamento é aplicada a toda coleção, entretanto somente em inglês. Assim, os documentos podem ser indexados por um motor de busca monolíngue. O sistema utilizado em todos os experimentos foi o *Zettair* (ZETTAIR, 2007).

A versão bilíngue (consultas em português sobre documentos em inglês neste exemplo) gerada com a técnica proposta é comparada com uma versão monolíngue (consultas e documentos em inglês), utilizando o motor de buscas *Zettair*, bem como com outras propostas enviadas ao CLEF nos mesmos anos utilizados para treinamento.

Além dos corpora acima citados, outro corpus foi utilizado. Este, entretanto, apenas para a geração de RAs, trata-se das transcrições de reuniões do parlamento europeu (EuroParl). Ele foi adicionado por ter sido traduzido manualmente, ao contrário dos anteriores, que são monolíngues.

```
<DOC>
<DOCNO>GH950103-000012</DOCNO>
<TEXT>
1. E=dan E=marino E=play E=home E=joe
   E=robbi E=stadium E=outplay E=joe
   E=montana E=saturday E=shoot E=between
   E=american E=footbal E=greatest
   E=quarterback E=threw
2. E=touchdown E=miami E=dolphin
   E=capitalis E=late E=turnov E=beat
   E=kansa E=city E=27 E=17 E=nfl E=play
   E=off
3. E=afc E=east E=champion E=miami E=visit
   E=afc E=west E=champion E=san E=diego
   E=first E=montana E=marino
4. E=contest E=1985 E=final E=marino
   E=complet E=22 E=29 E=pass E=257 E=yard
   E=montana E=hit E=26 E=37 E=pass
5. E=314 E=yard E=touchdown E=green E=bay
   E=packer E=contain E=leagu E=rush
   E=champion E=barry
6. E=sander E=held E=late E=detroit
   E=drive E=beat E=lion E=16 E=12 E=nfc
   E=wild E=card E=yesterday E=vinny
7. E=testaverd E=pass E=268 E=yard E=on
   E=touchdown E=lead E=cleveland E=brown
   E=past E=new E=england
8. E=patriot E=20 E=13 E=brown E=back
   E=testaverd E=defenc E=intercept E=new
   E=england E=quarterback
9. E=drew E=bledso E=three E=time
   E=cleveland E=meet E=archriv E=steeler
   E=saturday E=pittsburgh
</TEXT></DOC>
```

```
<DOC>
<DOCNO>GH950103-000012</DOCNO>
<TEXT>
   1. marin jog cas estadi joe robbi joe
      mont outplayed sab shoot out dool
      mai zagu futebol americ
   2. dool touchdowm miam dolphim capit
      dool turnov tard bat kans city 27
      17 nfl play off
   3. east campeao miam visit afc west
      campeao san dieg
   4. mont marin concurs final 1985 mar
      complet 22 29 pass 257 jard
   5. acert 26 37 pass 314 jard dool
      touchdowm
   6. bay pack cont campeao lig press
      barry sand realiz for tard carr bat
      detroit liom 16 12 cnf wild card
   7. vinny testav pass 268 jard
      touchdown lev cleveland browm pass
      new england patriot 20 13
   8. testav apoi com defes intercept nov
      inglaterr quarterback drew bledso
      tre vez
   9. reun archriv sab pittsburgh steel
</TEXT>
</DOC>
```

Figura 5.4: Documento traduzido com separação por frases.

5.2 Experimentos

Para realização dos experimentos foram analisados quatro aspectos, de forma a aferir a amplitude e eficiência da proposta. Assim, conforme citado no início do capítulo, foram variados: as consultas e o idioma das consultas, nesse caso, variando entre português e finlandês; o tipo de corpus utilizado na busca, que variou entre notícias de jornais e metadados bibliográficos; a forma de geração do corpus paralelo, que pode, com base no próprio corpus, em um corpus diferente, mas do mesmo domínio, e um corpus totalmente diferente, inclusive variando a forma de geração desse corpus de extração das RAs. Por fim, foi testada também a combinação da técnica proposta com técnica de tradução automática. Sempre que possível, foram feitas comparações com outros trabalhos semelhantes. Em todos os trabalhos, vale ressaltar que, como a proposta visa apenas resolver o problema do mapeamento de termos entre idiomas, é aceitável que o resultado não seja tão alto como os melhores do CLEF, uma vez que estes incluem técnicas de melhoria de desempenho, como realimentação de relevantes e expansão de consultas, não utilizadas neste trabalho.

Além de comparar com outros trabalhos, também apresenta o resultado de um sistema monolíngue executando-se os mesmos passos de pré-processamento e utilizando-se o mesmo SRI (*Zettair*). Além disso, as consultas foram editadas para adequarem-se à grafia do português brasileiro (por exemplo, “acção” foi convertida para “ação”), já que os tópicos do CLEF estavam na variante linguística de Portugal.

5.2.1 Variação do conjunto de consultas

Nesses experimentos foram avaliados os impactos da variação do conjunto de consultas. O corpus paralelo usado para a geração das regras foi mantido. Para este experimento foram utilizadas consultas das campanhas CLEF 2002 e 2005. Esses anos foram escolhidos por terem sido os únicos até 2007 com consultas em português (período em que o corpus trazia textos completos, ao contrário de 2008, quando passou a ser formado por metadados).

Para este experimento foi utilizado como corpus paralelo 20% das edições do *Jornal Los Angeles Times* de 1994 traduzidas automaticamente pelo *Google Translator*. O corpus de busca na versão de 2002 incluiu toda a coleção *LA Times*. Na versão de 2005 foram também utilizados na busca os documentos do *Jornal Glasgow Herald*, de 1995.

5.2.1.1 Tópicos Campanha CLEF 2002 multilíngue português/inglês

Para essa edição da Campanha CLEF foram comparados os sistemas abaixo:

- **Monolíngue** – Utilização do motor de busca *Zettair* para recuperação de documentos a partir das consultas em inglês, aplicando remoção de *stopwords* e remoção de sufixos usando o *Porter Stemmer* disponível no motor de buscas.
- **RAs** – Técnica básica de algoritmos de mineração de regras de associação conforme descrito no Capítulo 4, utilizando o algoritmo básico do *Zettair*.
- **RAs + Otimização** – Baseada na técnica de RAs, utilizando BM25+.
- **JHU/APL** (MCNAMEE; MAYFIELD, 2003) – Sistema de RI multilíngue, não dependente de idioma, baseado em n-gramas (n=6). Utiliza-se de um pequeno corpus artificial para traduções das consultas e realimentação de relevantes.
- **IRIT** (MOULINIER; MOLINA-SALGADO, 2003) – Traduz a consulta através de um tradutor *on-line* (Babelfish), e, para melhorar a busca monolíngue ou bilíngue, utiliza-se de um tesauro construído especificamente para o experimento e de um dicionário disponível na Internet.
- **TA** – tradução da consulta através de tradutor disponível na *Web* (*Google* tradutor), e aplicação do processo de tradução idêntico ao monolíngue.

Os resultados de 2002 em termos de precisão média dos testes acima citados são exibidos na Tabela 5.1. A utilização da métrica Okapi BM25+ se mostrou bastante importante para o resultado obtido se aproximar ao monolíngue.

Tabela 5.1: Comparativo com consultas e resultados do CLEF de 2002.

Experimento	Precisão média
Monolíngue	0,3718
RAs	0,1816
RAs + Otimização	0,3216
JHU/APL	0,4158
IRIT	0,2449
Tradução automática	0,3512

Fazendo-se um t-test sobre os dados, verificamos que a diferença entre os sistemas "Monolíngue" e "RAs + Otimização" não é estatisticamente significante (*p-value* = 0.36), o que demonstra a proximidade deste em relação ao resultado monolíngue. A Figura 5.5 traça as curvas de revocação *versus* precisão interpolada para os testes com os tópicos de 2002.

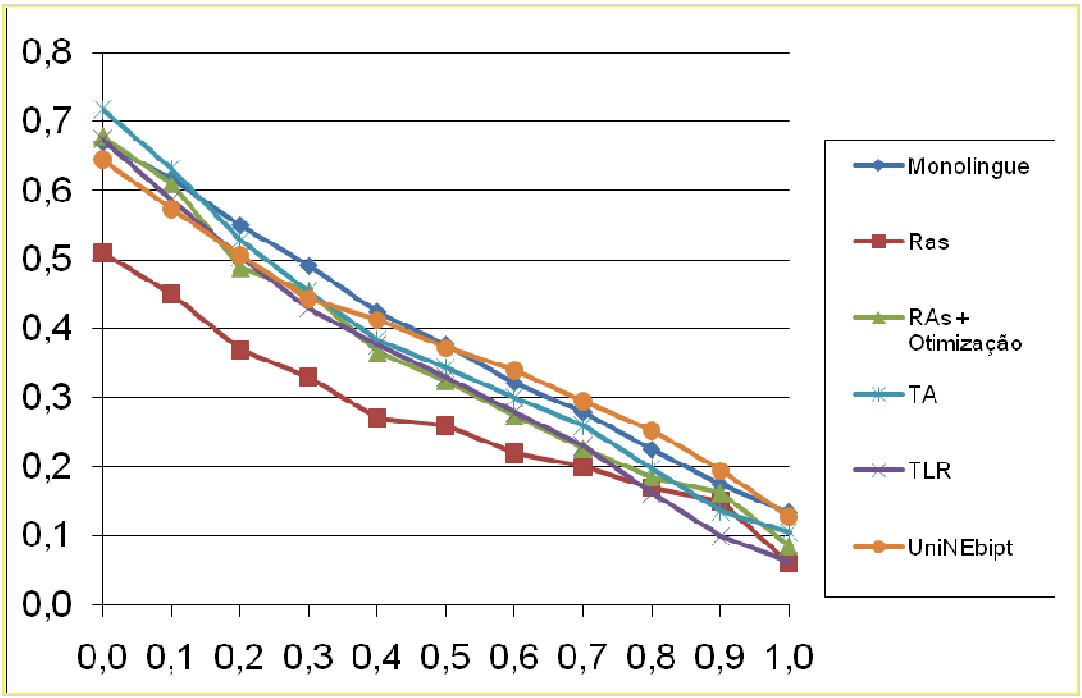

Figura 5.5: Precisão *versus* revocação 2002.

5.2.1.2 Tópicos Campanha CLEF 2005 multilíngue português/Inglês

Nesta edição da Campanha CLEF também foram disponibilizadas 50 consultas nos mais diversos idiomas (Búlgaro, Francês, Húngaro e Português), todas baseadas nas consultas monolíngues disponíveis em inglês. A coleção alvo da busca consistia de notícias dos jornais *Los Angeles Times de 1994* e *Glasgow Herald* do ano de 1995. A Tabela 5.2 relata os resultados dos experimentos realizados, análogos aos de 2002, e os compara com dois

trabalhos publicados no CLEF do mesmo ano. A Figura 5.6 exibe as curvas de precisão *versus* revocação do referido experimento.

Tabela 5.2: Comparativo CLEF 2005 com os documentos do *Los Angeles Times* e *Glasgow Herald*.

Experimento	Resultado
Monolíngue	0,3232
RAs	0,1829
RAs + Otimização	0,2469
Tradução automática	0,2929
TRL	0,2358

Fazendo-se um t-test sobre os dados, verificamos que a diferença entre os sistemas "Monolíngue" e "RAs + Otimização" não é estatisticamente significante (*p-value* = 0.052). O desempenho do método proposto pode ser considerado equivalente.

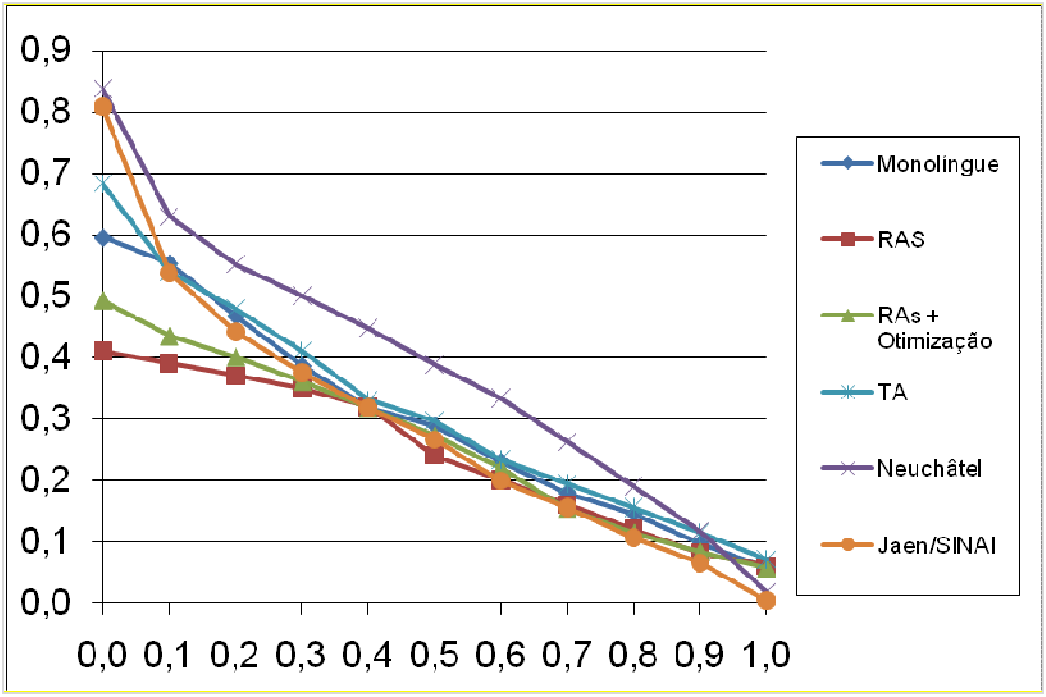

Figura 5.6: Precisão *versus* revocação 2005.

Traçando um comparativo com a versão monolíngue, obteve-se respectivamente 93,2% e 90,7% da MAP monolíngue, o que é considerado muito bom, visto que os melhores sistemas de RI-ML atingem entre 80 e 95%. Esse dado, somado aos t-tests que determinaram a não

significância estatística da perda em relação ao resultado monolíngue, atesta a eficiência do algoritmo diante de um total de 100 consultas.

5.2.2 Variação do idioma

Esse experimento visa determinar a eficiência da proposta em diferentes idiomas. Para tanto, utilizaram-se novamente as consultas de 2002 do CLEF. Um exemplo de consulta disponível em três idiomas (inglês, português e finlandês) é ilustrado Figura 5.7.

```
Inglês
<top>
<num>140</num>
<title>Mobile phones</title>
<desc>Prospects for the use of cellular phones.</desc>
</top>

Português
<top>
<num> 140 </num>
<title> Telefones móveis </title>
<desc> Perspectivas sobre o uso de telefones celulares. </desc>
</top>

Finlandês
<top>
<num> 140 </num>
<title> Matkapuhelimet (kännykät). </title>
<desc> Matkapuhelimien käytön ennusteet. </desc>
</top>
```

Figura 5.7: Tópico CLEF 2002 em diferentes idiomas.

A fim de testar a eficiência da técnica em diferentes idiomas, utilizaram-se corpora paralelos com os seguintes pares de idiomas: português-inglês e finlandês-inglês. A língua portuguesa foi escolhida por ser a língua nativa da autoria deste trabalho e também por ser bastante representativa em número de falantes nativos. Já a escolha da língua finlandesa se deve a sua grande diferenciação em relação ao português e ao inglês, não possuindo qualquer ancestral comum com essas línguas, apesar de utilizar o mesmo alfabeto. Nesse sentido, a fim de validar o impacto do idioma das consultas no método descrito na seção 5.1.4, como corpus paralelo foi utilizado o corpus *Los Angeles Times* sendo referenciados por **RA-LATimes**. Conforme descrito na seção 5.1.4, a escolha dos documentos para tradução é feita sem levar em consideração a sua relevância para as consultas.

Para se ter uma base de comparação que nos permita avaliar o quanto se perde nessa abordagem de RI-ML utilizando RAs, também se executou uma recuperação monolíngue. Esse *baseline* foi chamado de **Mono**.

Para todos os experimentos foram utilizados remoção de *stopwords* e *stemming*. Esses processos, a exemplo do que ocorreu no experimento anterior, são executados antes da paralelização do corpus, pois são dependentes de idioma. Assim, nesses experimentos foram utilizados o *Porter Stemmer* (PORTER, 1980) para o corpus de busca em inglês e para a parte inglesa dos corpora paralelos, e sua variante para a língua finlandesa para a partição nesse

idioma do corpus de mineração. Já para a língua portuguesa foi utilizado o removedor de sufixos da língua portuguesa (COELHO; ORENGO; BURIOL, 2007).

Após a obtenção dos corpora paralelos de busca e de extração de regras de associação, se executa o algoritmo de RAs proposto na seção 4.4, a fim de identificar os termos em inglês equivalentes aos termos em finlandês e português de cada consulta. Um exemplo de tópico é demonstrado na Figura 5.8.

```
RA-LATimes Português
<top>
<num>140 </num>
<title> phone mobil</title>
<desc> perspect phone telephon cell mobil</desc>
</top>
RA-LATimes Finlandês
<top>
<num> 140 </num>
<title> cellular mobil perspect  </title>
<desc> perspect cellular cellular telephon </desc>
```

Figura 5.8: Exemplo de consulta processada.

Após obtenção das consultas devidamente traduzidas para a língua inglesa, elas são submetidas ao motor de RI monolíngue *Zettair*. Vale salientar que o método adotado é totalmente automático, e seguiu estritamente a abordagem descrita na seção 4.4, não havendo qualquer tipo de intervenção humana na seleção dos termos. E também não foram executadas quaisquer realimentações de relevantes, pseudorrelevantes ou expansão de consultas. Entretanto, a seleção de mais de uma possível tradução pode ser considerada por algumas bibliografias como uma expansão de consultas implícita.

5.2.2.1 Resultados com Variação do Idioma

Os resultados dos experimentos são demonstrados na Figura 5.9 e na Tabela 5.3. A Figura 5.9 demonstra as curvas precisão *versus* revocação para cada um dos testes. A Tabela 5.3 contém os valores de MAP e duas outras estatísticas úteis: o número de termos utilizados na busca e o número de termos que não foram possíveis de mapear utilizando a técnica proposta.

Os experimentos que utilizaram algoritmos de RAs para mapeamento de termos atingiram até 88% do desempenho do monolíngue, que é comparável ao estado da arte. Os melhores resultados foram obtidos utilizando uma amostra do corpus de consultas para a mineração de RAs. O teste-t demonstrou não haver diferença significativa entre a execução monolíngue e as execuções utilizando RAs em português e finlandês (*p-value* de 0,08 e 0,06, respectivamente). Os resultados também demonstram que o desempenho do experimento RA-*LATimes* é consistente em ambos os idiomas, isso é exemplificado pela alta correlação entre os documentos retornados em um e outro idioma.

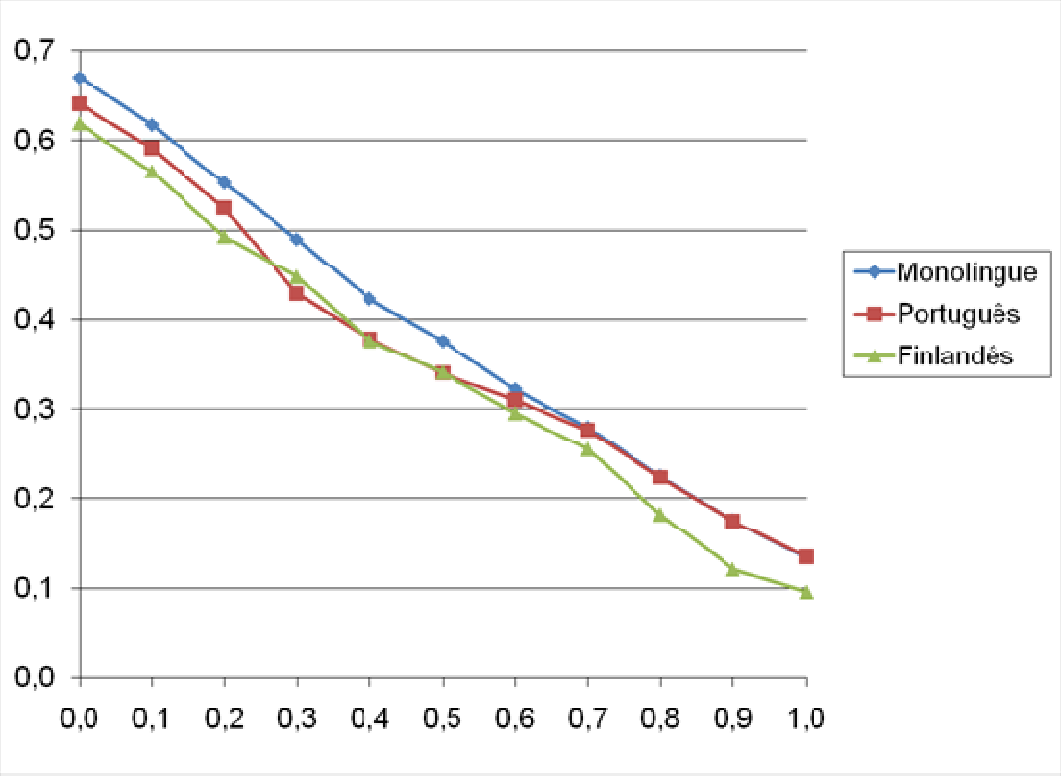

Figura 5.9: Precisão *versus* Revocação para Consultas CLEF 2002 em português e finlandês.

Tabela 5.3: Resultados da variação do idioma das consultas.

Experimento	Língua	MAP	Termos traduzidos	Termos não traduzidos
Mono	Inglês	0.4423	946	—
RA-*LATimes*	Português	0.3786	848	64
RA-*LATimes*	Finlandês	0.3895	650	103

Para realizar uma comparação justa entre essa proposta e outros grupos, é necessário que eles tenham utilizado o mesmo conjunto de questões, sobre a mesma coleção e na mesma língua. Assim, levantou-se apenas dois estudos que preenchem esses requisitos: Orengo e Huyck (2003) e McNamee e Mayfield (2003). Desses RAs, superam significativamente a primeira abordagem baseada em Indexação semântica latente (DEERWESTER, 1990), a qual também utilizou um sistema de TA para simular um corpus paralelo e obteve MAP = 0,2088 que representava 87% do equivalente monolíngue. Os resultados reportados em McNamee e Mayfield (2003) são superiores aos obtidos neste trabalho (MAP = 0,4158), entretanto este utilizou expansão de consultas para até mesmo superar o resultado monolíngue.

Acredita-se que a razão para a maior queda no desempenho, quando esta ocorre, tenha sido a tradução errada. Por exemplo, na consulta 114, a palavra “líder”, a qual tem como

tradução "*leader*", foi erroneamente traduzida para "*drive*". Acrônimos também são um problema na abordagem proposta. Seu equivalente em outra língua nem sempre é encontrado pelas RAs, e quando o eram, em sua forma expandida nem sempre todos os termos eram utilizados, pois o reconhecimento de multipalavras é falho na proposta. Outra razão para o mau desempenho foi a falta de traduções para alguns termos. Isso ocorre especialmente com entidades nomeadas como "*Eurofighter*" (tópico 93) e "Ames" (tópico 100). Uma tentativa de identificação e tratamento dessas entidades foi executada, mas como seus resultados acabaram prejudicando os resultados das demais consultas, esta foi descartada, ficando para um trabalho futuro.

5.2.3 Variação do corpus de geração das RAs

Esse experimento visa determinar a eficiência da proposta em diferentes corpora paralelos para geração de RAs, sejam estes corpus traduzidos artificialmente ou manualmente. O conjunto de consultas utilizado é o mesmo da seção 5.2.2, entretanto foram adicionadas variantes no corpus utilizado para extração das RAs mantendo-se os mesmos pares de idiomas. Assim, foram utilizados os seguintes *corpora*, além dos utilizados anteriormente:

- Utilizando corpora distintos para extração das RAs e para execução das consultas. Para tanto, se traduziu automaticamente uma amostra de 20% dos documentos do *Jornal Glasgow Herald* também para o português e para o finlandês, gerando, dessa forma, um corpus paralelo sintético. Esse corpus, apesar de distinto, possui a mesma temática do corpus utilizado nas consultas. Esse experimento será referenciado como **RA-GH**.
- Utilizando um corpus de domínio diferente para geração das RAs. A ideia é testar a viabilidade da utilização de um corpus paralelo com tradução de boa qualidade (traduzido manualmente) para obtenção das RAs. O corpus utilizado é composto de discussões do parlamento europeu. Esses discursos são traduzidos manualmente para todas as línguas da União Europeia, o que compreende o português, o finlandês, entre outras. Para tanto, foram paralelizados os documentos, e, como esta coleção utiliza-se dos documentos referenciados como *EuroParl*, o experimento será referenciado como **RA-*EuroParl***.
- A fim de testar se RAs e MT poderiam ser utilizados em conjunto para melhorar os resultados, testamos também a combinação dos melhores RAs com os melhores TA executados. A combinação foi feita por meio da realização de um conjunto união dos termos de consulta gerados a partir de duas estratégias. Essas execuções são chamadas **TA + AR**.

Com o objetivo de ter alguns meios de comparação entre a abordagem AR-MT e de outras abordagens, traduzimos também apenas os tópicos das consultas para português e finlandês partindo das consultas originais em inglês. Para isso utilizou-se o *Google* Tradutor e o *LEC Power Translator* (LEC..., 2006). Entretanto, o *LEC Power Translator* não possui suporte para finlandês, neste caso utilizou-se apenas para o português. Após a tradução dos tópicos, procedemos uma recuperação monolíngue. Esse experimento foi chamado de **TA-*Google*** e **TA-*LEC***.

De forma análoga, as etapas de pré-processamento foram executadas com remoção de *stopwords*, de sufixos e pontuações conforme descrito na seção 4.4.1 a fim de identificar os termos em inglês equivalentes aos termos em finlandês e português de cada consulta. Um do mesmo tópico citado na seção 5.2.2 é mostrado na Figura 5.10.

```
RA-GW Português
<num>140 </num>
<title> phone mobil</title>
<desc> perspect phone cell</desc>
RA-EuroParl Português
<num>140 </num>
<title> telephon mobil</title>
<desc> perspect telephon cell</desc>
RA-GW Finlandês
<num> 140 </num>
<title> cellular  mobil  perspect   </title>
<desc> perspect   cellular  cellular telephone  </desc>
RA-EuroParl Finlandês
<num> 140 </num>
<title> cellular  mobil  perspect   </title>
<desc> perspect   cellular  cellular  </desc>
TA-Google Português
<num> 140 </num>
<title> Mobile phones </title>
<desc> perspective on the use of mobile phones </ desc>
MT-LEC
<num> 140 </num>
<title> movable Telephones </title>
<desc> Perspectives on the use of cellular telephones. </desc>
RA-TA-Google Finlandês
<num> 140 </num>
<title> Cell phones (mobile phones). </title>
<desc> of cell phones using the forecasts. </desc>
```

Figura 5.10: Exemplo de consulta processada.

5.2.3.1 Resultados com Variação do corpus *paralelo*

Os resultados dos experimentos são demonstrados nas Figura 5.11 e 5.12 e na Tabela 5.3. Nesses resultados foram adicionados, além dos resultados aqui produzidos, a versão monolíngue das consultas como forma de comparação. A Tabela 5.3 contém os valores de MAP e duas outras estatísticas úteis: o número de termos utilizados na busca e o número de termos que não foram possíveis de mapear utilizando a técnica proposta.

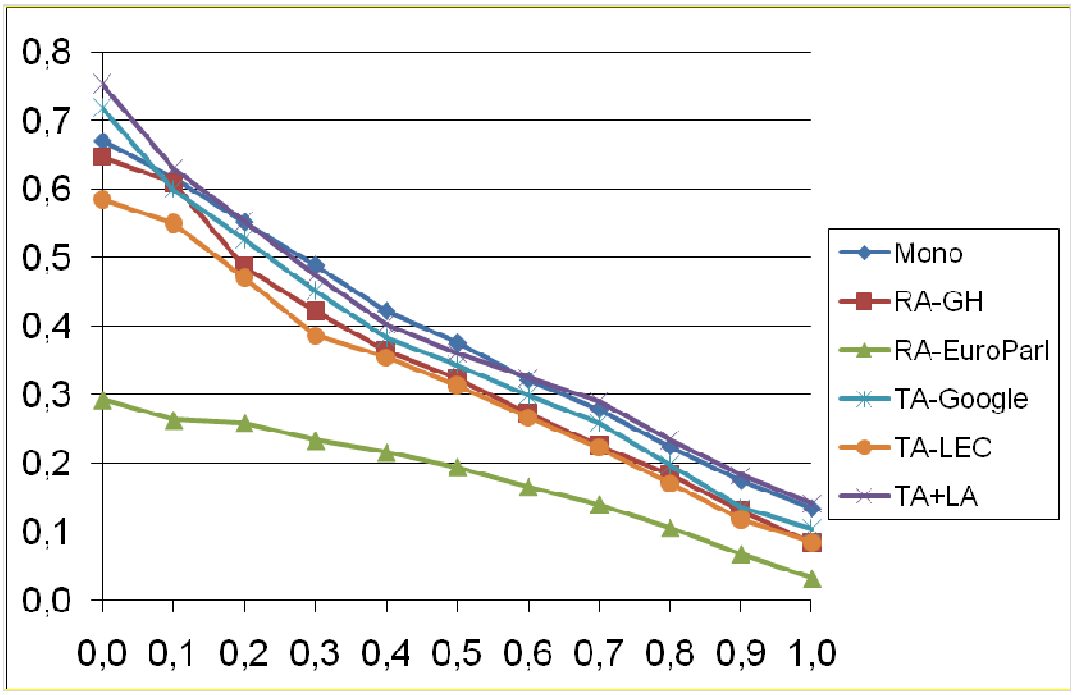

Figura 5.11: Precisão *versus* revocação para consultas CLEF 2002 em português.

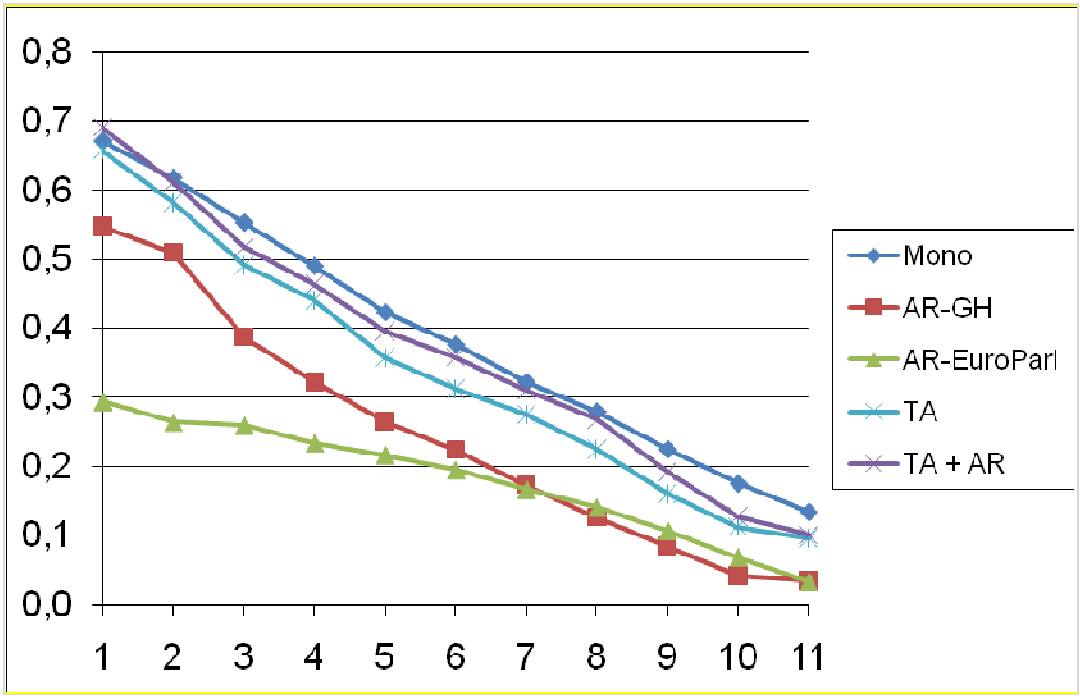

Figura 5.12: Precisão *versus* revocação para consultas CLEF 2002 em finlandês.

Os experimentos que utilizaram algoritmos de RAs com um corpus diferente ao da coleção atingiram 83% da versão monolíngue (baseado na utilização do *Glasgow Herald*), contra 88% quando da utilização da mesma coleção. Da mesma forma que no teste anterior, o teste-t demonstrou não haver diferença significativa entre a execução monolíngue e as execuções utilizando RAs em português e finlandês. (*p-value* de 0,054 e 0,09, respectivamente). Já os experimentos utilizando o corpus de pronuncionamentos do congresso europeu, AR-*EuroParl* foi significativamente inferior à recuperação monolíngue (*p-value* de 0,0004 e 0,0001, respectivamente). A Tabela 5.4 sumariza os resultados e os compara com a versão monolíngue.

Tabela 5.4: Experimentos multicorpora.

Experimento	Língua	MAP	Termos traduzidos	Termos não traduzidos
Mono	Inglês	0.4423	946	—
RA-GH	Português	0.3317	869	80
RA-*EuroParl*	Português	0.1954	632	67
RA-*LATimes*	Português	0.3786	848	64
TA-LEC	Português	0.3659	1047	—
TA+RA	Português	0.4942	1064	—
RA-GH	Finlandês	0.3646	628	118
RA-*EuroParl*	Finlandês	0.2790	497	21
TA-*Google*	Finlandês	0.3782	868	—
TA+RA	Finlandês	0.3981	910	—

A Tabela é análoga a utilizada na seção 5.5.2. A junção de métodos de TA com RAs mostrou bons resultados em ambos os idiomas, acredita-se que isso se deva ao fato desta junção combinar com a maior quantidade de termos com traduções conseguidas por métodos de tradução automática, aliada à possibilidade de obtenção de múltiplas traduções com seus respectivos pesos obtida pelo método proposto baseado em RAs.

Os experimentos baseados em TA utilizando o *Google Translate* superaram RAs para as consultas em português, mas não o fizeram para o finlandês. Entretanto, em ambos os casos, não houve diferença estatística com o experimento RA-*LATimes*; obtendo RA-MT para português ($p = 0,36$) e para o finlandês ($p = 0,58$). Embora com resultados não tão bons, também foi executada uma trilha utilizando outro tradutor automático o *LEC Power Translator* (TRANSLATE, 2006), mas este não está disponível para a língua finlandesa e obteve resultados significativamente inferiores. Apesar de a princípio ser esperado um desempenho muito melhor de MT, devido ao emprego de métodos muito mais sofisticados de processamento de linguagem natural, na prática isso não se confirmou, em parte pela seleção

de múltiplos termos utilizada pela proposta. Como a nossa abordagem é muito simples e se baseia apenas em estatísticas obtidas a partir de um corpus paralelo, essa falta de diferença significativa favorece a simplicidade computacional e independência de idioma dessa proposta.

A partir de uma análise tópico por tópico, foi possível observar que algumas consultas foram ajudadas por essa abordagem e outras prejudicadas. Uma das tendências é a obtenção de melhores resultados com TA quando há grande número de termos nas consultas.

Por outro lado, a abordagem baseada em RAs foi melhor que TA e até mesmo que a versão monolíngue em alguns tópicos, especialmente pela adoção de mais de uma tradução para o mesmo termo. Por exemplo, no tópico 122 o termo "internacional" foi traduzido para "international" e "global". Com isso, o resultado de MAP dessa consulta superou em 426% o resultado TA-*Google* e em 288% ao resultado monolíngue.

5.2.4 Variação do corpus de busca

A fim de comparar a técnica proposta com outras técnicas sobre um corpus totalmente novo, submeteu-se o algoritmo proposto neste trabalho à campanha de avaliação CLEF de 2008 na categoria de recuperação de informação bilíngue, utilizando, para isso, consultas em espanhol também sobre um corpus em inglês, porém, dessa vez, o corpus não se encontrava totalmente em inglês, já que havia livros com títulos em outros idiomas, mas essa informação não foi considerada no experimento. Essa posição também foi adotada pelos demais trabalhos submetidos à conferência de avaliação (PETERS, 2009), como forma de também testar a eficiência do experimento em um terceiro idioma.

As consultas da campanha CLEF de 2008 diferem radicalmente das anteriores, tendo sido substituídos o corpus baseado em notícias de jornais em SGML com registros extensos de em média 2 kb em seu campo de corpo de notícia, por um formado por metadados em XML de informações bibliográficas da Biblioteca Inglesa (TEL) com campos de em média 6 termos. A Tabela 5.5 detalha as características da coleção.

Tabela 5.5: Características corpus CLEF 2008.

Número de termos únicos	689.053
Número de documentos	1.000.101
Tamanho	195 MB

O procedimento é o mesmo descrito na seção 4.4. Assim, a nossa abordagem necessita de uma amostra de documentos paralelos da coleção TEL. Como esta não tem documentos paralelos, extraímos uma amostra de 25% (250.025) da mesma e efetuamos o pré-processamento para alinhamento como descrito na seção 4.4.1. A amostra foi obtida extraindo-se um a cada 4 documentos da coleção. Utilizamos a lista de *stopwords* e o *stemmer Porter* em suas versões para espanhol e inglês. As etapas de pós-processamento foram análogas aos experimentos anteriores utilizados. Entretanto, devido a maior quantidade de documentos, o tempo de processamento foi de 45 segundos, incluindo a extração dos RAs, regra de filtragem, consulta de tradução e processamento pelo motor de pesquisa, utilizando os equipamentos detalhados na seção 5.1.2.

A submissão à conferência constou de quatro experimentos: dois bilíngues espanhol/inglês e dois monolíngues que serviram de *baseline*. Os experimentos monolíngues objetivaram testar as diferenças de performance dos algoritmos BM25 e BM25+, enquanto as consultas bilíngues, além destas, também testaram a heurística de mapeamento entre o par de idiomas, a qual é o cerne deste trabalho. Dessa forma foram submetidos:

- **RAs** – Utilizando a proposta deste trabalho e o algoritmo de RI BM25;
- **RAs BM25+** – Utilizando a proposta deste trabalho e o algoritmo de RI BM25+;
- **Mono** – monolíngue utilizando o algoritmo BM25;
- **Mono BM25+** – monolíngue utilizando o algoritmo BM25+.

Os resultados obtidos por este trabalho são resumidos na Tabela 5.6 e na Figura 5.13. Comparando os experimentos monolíngues e bilíngues percebemos que a execução bilíngue atingiu 86% do correspondente monolíngue em termos da MAP em linha com os experimentos anteriores. Um t-test mostrou que a diferença de desempenho entre a versão monolíngue e bilíngue não é estatisticamente significante em termos da MAP. Isso foi notado tanto para a roda com a versão original do BM25 quanto para a versão com o BM25+.

Comparado aos outros participantes, a nossa versão bilíngue foi classificada em terceiro lugar. Esses resultados indicam que a abordagem para o mapeamento de conceitos entre idiomas utilizando RAs é adequada. Os trabalhos desses demais participantes são descritos brevemente na seção 3.2.

Comparando-se também os resultados obtidos com a versão original do algoritmo Okapi BM25 com a versão BM25+, nota-se que estas são estatisticamente significantes em termos de MAP.

Tabela 5.6: Experimentos CLEF 2008.

Experimento	Precisão média
Monolíngue	0.2493
Monolíngue BM25+	0.2777
RAs	0.2151
RAs BM25+	0.2315

A Figura 5.14 demonstra os valores de precisão deste trabalho em comparação com os demais cinco primeiros colocados na submissão 2008 da campanha CLEF (PETERS, 2009). Também foi submetido nessa mesma conferência outro artigo utilizando esta proposta em conjunto com reconhecimento de multipalavras, obtendo primeiro lugar na respectiva categoria, a qual utiliza informações de desambiguação de termos.

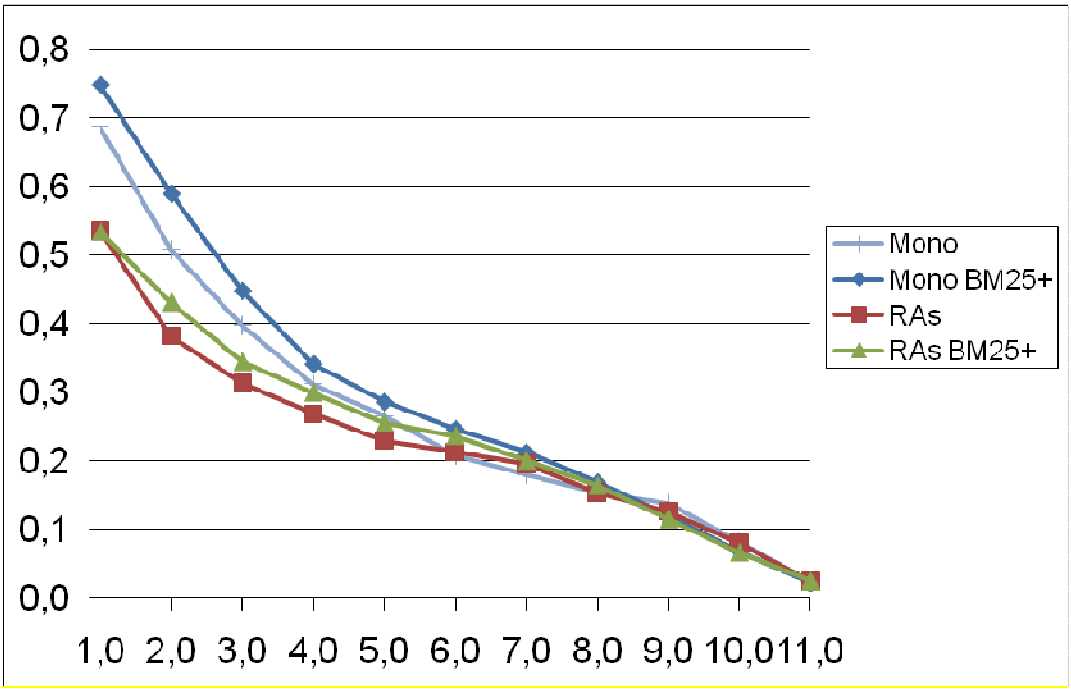

Figura 5.13: Curva de precisão *versus* revocação CLEF 2008.

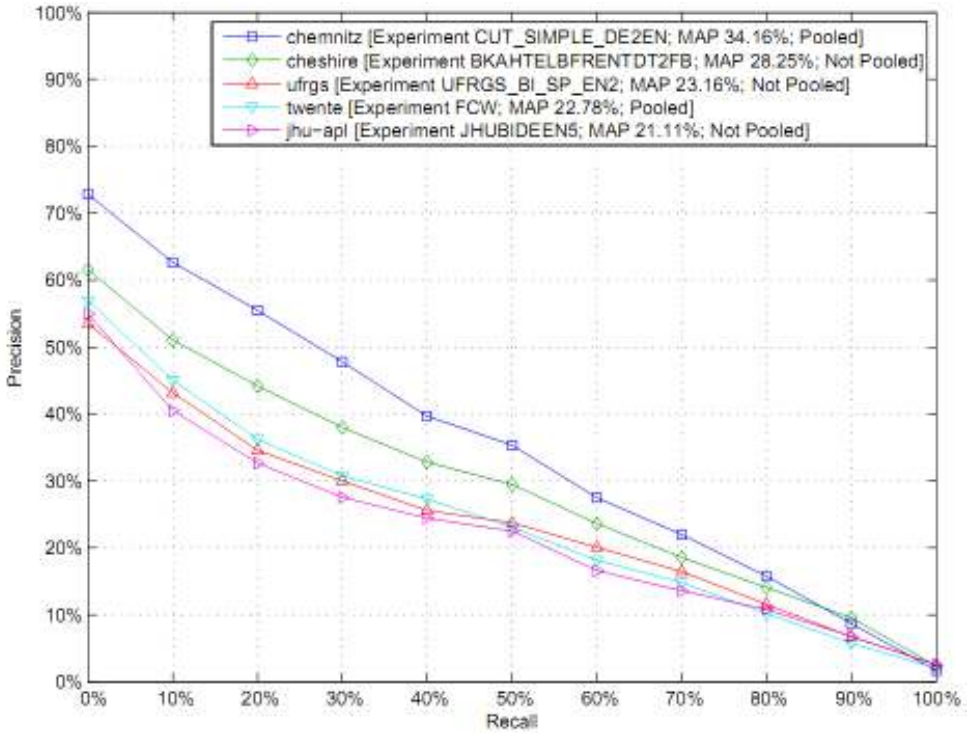

Figura 5.14: Resultado oficial CLEF 2008 (CLEF, 2009).

Os resultados obtidos são particularmente bons se considerarmos que não foram utilizados realimentação de relevantes nem qualquer tipo de tesauro como os trabalhos concorrentes, os quais são analisados na seção 3.2.

5.3 Conclusões

Com base no *set* de experimentos executados, concluiu-se que em todos os resultados houve uma perda em relação ao monolíngue insignificante no melhor caso; que este se aplica com boa qualidade de resposta desde documentos até metadados; e a variação do idioma também se mostrou pouco impactante no resultado. Entretanto, a utilização de corpus com tradução manual mostrou-se bastante impactante no resultado, mas acreditamos que isso se deva pela especificidade do corpus natural utilizado, baseado em temas políticos apenas. Infelizmente não foi possível utilizar outro corpus natural pela dificuldade de acesso a eles, sobretudo em línguas com pouca quantidade de fluentes, como o finlandês.

Acreditamos na eficiência da técnica principalmente entre idiomas nos quais não há ferramentas de tradução direta. Como a maioria das ferramentas se resume a poucas línguas e utiliza um pivô de tradução, a técnica aqui proposta existindo um corpus paralelo entre o par de línguas utilizadas não necessita desse pivô.

O próximo capítulo demonstra brevemente o protótipo construído e utilizado para demonstrar essa técnica por meio da Internet.

6 PROTÓTIPO PARA CONSULTAS À *WEB*

Este capítulo apresenta um protótipo para RI-ML que aceita consultas (palavras-chave), as traduz para inglês e as envia para diversos motores de busca na *Web*. A consulta traduzida irá recuperar documentos textuais, imagens e vídeo. Por meio do protótipo, usuários da *Web* que falam apenas português poderão ter acesso a um universo muito maior de informações.

O protótipo, chamado *ConsultaToSearch*, está acessível a partir do endereço http://www.inf.ufrgs.br/~apgeraldo/busca. A estratégia para a mineração das RAs é *eager* (seção 4.4.5), uma vez que o tempo resposta baixo é primordial na Internet.

Para a seleção dos termos que comporiam a cache do *ConsultaToSearch* foram contadas as ocorrências de todos os termos do Jornal Folha de São Paulo de 1994 e selecionados todos os termos com 10 ou mais ocorrências, resultando em um total (excluindo as *stopwords*) de aproximadamente 34.000 termos. Foram extraídas as traduções para esses termos usando o método descrito na seção 4.4. O resultado desse processo, isto é, os termos em português e suas traduções, foi então armazenado para servir de dicionário eletrônico para o *ConsultaToSearch*.

O funcionamento é bastante simples. O usuário apenas deve escrever sua consulta utilizando palavras-chave em português. Essa consulta é traduzida, substituindo-se cada termo por suas traduções e a consulta resultante é submetida a uma série de *sites* de busca textual como:

- Buscadores gerais (*Google*, *Yahoo*, *Live*, *Lycos*);
- Bibliotecas de imagens (*Google* Imagens, *Yahoo* Imagens);
- *Sites* de vídeos (*YouTube*, *Yahoo* Videos);
- Bibliotecas digitais (*ACM*, *Scholar Google*);
- Lojas *online* (*E-bay*, *Amazon*);
- Enciclopédias (Wikipédia).

A Figura 6.1 exemplifica uma consulta na aplicação. A fim de auxiliar o usuário na digitação de termos, as palavras existentes no dicionário bilíngue são sugeridas à medida que o usuário digita. Também é possível adicionar novos *sites* por meio de uma opção disponível no final da lista de *sites*. Dessa forma, *ConsultaToSearch*, além de permitir a busca multilíngue, também pode ser usada como uma forma de aceleração de buscas, uma vez que as consultas são sempre executas em paralelo através de chamadas assíncronas aos diversos *sites*.

Nos casos em que termos ambíguos são usados na consulta, *ConsultaToSearch* mostra um par de documentos para o usuário, que decide qual é mais próximo de seu objetivo, assim escolhendo qual termo será utilizado na busca. A Figura 6.2 exemplifica a busca por “bateria descarregando”. A palavra bateria pode ser traduzida para língua inglesa como *battery* ou

drums. A ferramenta exibe um resultado com cada uma das traduções candidatas para que o usuário faça a seleção.

O protótipo foi implementado utilizando a linguagem PHP, que é amplamente disponível em servidores HTTP. As páginas geradas possuem carregamento de informações por *Java script*, assim permitindo a execução assíncrona das solicitações, bem como evitando novas solicitações de dados já consultados na mesma execução.

Figura 6.1: Exemplo de consulta RI-ML prorocessada pelo protótipo.

Figura 6.2: Desambiguação de termos.

Existem ainda possibilidades de melhorias, que incluem a utilização de um dicionário mais abrangente e adição de dicionários de domínios específicos.

7 CONCLUSÃO

O presente trabalho apresenta um método de RI-ML para mapeamento de termos entre diferentes idiomas, utilizando corpora paralelos multilíngues. A proposta pode ser aplicada tanto a sistemas de busca sobre coleções indexadas localmente, bem como sobre a Internet. Para tanto, este trabalho propõe uma série de etapas necessárias à utilização de algoritmos de mineração de RAs, permitindo, dessa forma, a filtragem das possíveis traduções de um termo.

Entre as principais contribuições deste trabalho, destacam-se:

- a proposta de um novo método para RI-ML. Os resultados obtidos pelo método proposto são comparáveis aos de um sistema monolíngue e aos resultados do estado da arte;
- a recuperação de documentos em um idioma em resposta a uma consulta formulada em outro idioma, dessa forma, permitindo a busca multilíngue;
- permitir a busca multilíngue de objetos gráficos com metadados tais como fotos, imagens e vídeos;
- definição de heurísticas para a filtragem de regras de associação que mais provavelmente mapeiam um termo a sua tradução, permitindo resultados melhores;
- desenvolvimento de um protótipo para consulta multilíngue sobre a *Web* que recebe uma consulta em português, a traduz para inglês usando o método proposto, e envia a consulta para diversos motores de busca.

Os experimentos testaram diferentes combinações de idiomas, diferentes coleções de dados, diferentes conjuntos de consultas e diferentes alternativas para a geração do corpus paralelo. Grande parte desses experimentos foi submetida a conferências e campanhas de avaliação internacionais. Na campanha de avaliação CLEF 2008, comparada aos demais trabalhos descritos no Capítulo 3, a proposta obteve terceiro lugar executando consultas sobre o catálogo de metadados bibliográficos da biblioteca inglesa. Nos testes utilizando as coleções de jornais *Los Angeles Times* e *Glasgow Herald*, o método proposto ficou em primeiro lugar. Além disso, testes estatísticos com os resultados dos experimentos mostram que não há diferença significativa em relação a consultas monolíngues.

Como produção científica foram publicados os seguintes artigos:

- Geraldo, A.P. and V.M. Orengo, UFRGS@CLEF2008: using association rules for cross-language information retrieval. In: Cross-Language Evaluation Forum, 9., 2008, Aarhus, Dinamarca. **Proceeding...** Heidelberg: Springer, 2009. p.66-74.

  Descreve o método proposto, bem como os experimentos realizados para a campanha CLEF 2008 com os metadados da biblioteca britânica.

- ACOSTA, O. C. et al. UFRGS@CLEF2008: Indexing Multiword Expressions for Information Retrieval. In: Cross-Language Evaluation Forum, 9., 2008, Aarhus, Dinamarca. **Proceeding...** Heidelberg: Springer, 2009.

  Aplicação do método proposto sobre coleções de jornais acima citados. Foi usado um extenso conjunto de 150 consultas.

- GERALDO, A. P.; MOREIRA, V. P.; GONÇALVES, M. A. On-demand associative cross-language information retrieval. In: International Symposium on String Processing and Information Retrieval, 16., 2009, Saariselkä, Finland. **Proceeding…** Heidelberg: Springer-Verlag, 2009. p.165-173.

  Trata da utilização de diferentes idiomas nas consultas, da variação de corpus utilizado para extração das RAs. Também compara os resultados das RAs com métodos de tradução automática e testa a combinação dos dois.

Dois outros trabalhos foram desenvolvidos em paralelo com esta dissertação também acerca de RI-ML e RI:

- GERALDO, A. P.; ORENGO, V. M. Ajustando a importância dos termos: uma extensão à BM25. In: Simpósio Brasileiro de Banco de Dados, 23., 2008, Campinas. **Anais...** Campinas: SBC, 2008.

  Define uma extensão à métrica BM25, que foi utilizada nos experimentos posteriores.

- FLEMMINGS, R. et al. BBK-UFRGS@CLEF2009: query expansion of geographic place names. In: Cross-Language Evaluation Forum, 2009, Corfu, Grécia. **Proceeding…** Heidelberg: Springer, 2009.

  Propõe uma estratégia para a expansão de consultas com termos geográficos.

Apesar dos bons resultados obtidos pelos experimentos, ainda há espaço para muitas melhorias. Neste trabalho foram apenas considerados os casos em que um termo no idioma origem é traduzido para um ou mais termos simples no idioma destino, não sendo considerados termos compostos em quaisquer dos idiomas, uma vez que não é rara a tradução de um termo para mais de um termo em outro idioma (por exemplo, saboneteira é traduzido para *soap dish*). Esses casos podem ser tratados com RAs adaptando-se as heurísticas de filtragem. Outros trabalhos, como Hull (1993) e Grefenstette (1998), indicam que traduzir frases em vez de simples palavras gera melhorias significativas.

Além dessas melhorias, a união de métodos linguísticos com as heurísticas puramente estatísticas propostas neste trabalho poderiam dar melhores resultados, especialmente quando são encontradas diversas traduções para um mesmo termo. Métodos de desambiguação também poderiam trazer benefícios.

Clustering de documentos poderia melhorar significativamente a escolha de traduções para os termos, desde que esta possa ser aplicada tanto à consulta quanto ao conjunto de documentos.

Por fim, algoritmos de realimentação de relevantes podem ser utilizados para definir a melhor tradução para um termo em um dado contexto.

REFERÊNCIAS

ACOSTA, O. C. et al. UFRGS@CLEF2008: Indexing Multiword Expressions for Information Retrieval. In: Cross-Language Evaluation Forum, 9., 2008, Aarhus, Dinamarca. **Proceeding...** Heidelberg: Springer, 2009.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: VLDB Conference, 1994, Santiago, Chile. **Proceedings…** Santiago, Chile: Morgan Kaufmann, 1994.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. Harlow, Reino Unido: Pearson, 1999. 513p.

BRUHA, I.; FAMILI, A. Postprocessing in machine learning and data mining. **SIGKDD Explor. Newsl.**, v.2, n.2, p.110-114, 2000.

CARDOSO, O. N. P. Recuperação de informação. **Journal of Computer Science**, v.2, n.1, 2000.

CHENG, P.-J. et al. Translating unknown queries with web corpora for cross-language information retrieval. In: Annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 27., 2004, Sheffield, United Kingdom. **Proceedings…** New York, USA: ACM, 2004. p.146-153.

CLINCHANT, S.; RENDERS, J.-M. XRCE's participation to CLEF 2008 ad-hoc track. In: Cross-Language Evaluation Forum, 9., 2008, Aarhus, Dinamarca. **Proceeding...** Heidelberg: Springer, 2009.

COELHO, A. R.; ORENGO, V. M.; BURIOL, L. RSLP: uma ferramenta para a remoção de sufixos na língua portuguesa. In: Simpósio Brasileiro de Banco de Dados, 22., 2007, João Pessoa, PB. **Anais…** João Pessoa, PB: SBBD, 2007. p.4-46.

COMUNIDADE DOS PAÍSES DE LÍNGUA PORTUGUESA (CPLP). **Ratificação da declaração constitutiva e dos estatutos da comunidade dos países de língua portuguesa**. Lisboa: CPLP, 2007. Disponível em: <http://www.cplp.org/docs/documentacao/Ratifica%C3%A7%C3%A3o%20da%20Declara%C3%A7%C3%A3o%20Constitutiva%20e%20dos%20Estatutos%20da%20Comunidade%20dos%20Pa%C3%ADses%20de%20L%C3%ADngua%20Portuguesa.pdf>. Acesso em: 20 nov. 2008.

COOPER, W. S.; GEY, F. C.; DABNEY, D. P. Probabilistic retrieval based on staged logistic regression. In: Annual international ACM SIGIR conference on Research and development in information retrieval, 15., 1992, Copenhagen, Denmark. **Proceeding…** New York, USA: ACM, 1992. p.198-210.

CROSS-LANGUAGE EVALUATION FORUM (CLEF). Desenvolvido por CLEF. Disponível em: <http://www.clef-campaign.org>. Acesso em: 18 set. 2007.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the Américan Society for Information Science**, v.41, p.1-13, 1990.

FAVARO, T.; VIEIRA V. Não será por falta de memória. **Revista Veja**, 30 abr. 2008.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v.39, n.11, p.27-34, 1996.

FAYYAD, U.; UTHURUSAMY, R. Data mining and knowledge discovery in databases. **Communication ACM**, v.39, n.11, p.24-26, 1996.

FIGUEROLA, C. G. et al. Stemming in Spanish: a first approach to its impact on information retrieval. In: Cross-Language System Evaluation Campaign, 2001, Darmstadt, Germany. **Results…** Darmstadt, Germany: IEI-CNR, 2001. p.197-202.

FLEMMINGS, R. et al. BBK-UFRGS@CLEF2009: query expansion of geographic place names. In: Cross-Language Evaluation Forum, 2009, Corfu, Grécia. **Proceeding…** Heidelberg: Springer, 2009.

FREDKIN, E. Trie memory. **Communication ACM**, v.3, n.9, p.490-499, 1960.

FUJII, A.; ISHIKAWA, T. Japanese/English cross-language information retrieval: exploration of query translation and transliteration. **Computers and the Humanities**, v.35, n.4, p.389-420, Nov. 2001, 2004.

GERALDO, A. P.; ORENGO, V. M. Ajustando a importância dos termos: uma extensão à BM25. In: Simpósio Brasileiro de Banco de Dados, 23., 2008, Campinas. **Anais...** Campinas: SBC, 2008.

GERALDO, A. P.; MOREIRA, V. P. UFRGS@CLEF2008: using association rules for cross-language information retrieval. In: Cross-Language Evaluation Forum, 9., 2008, Aarhus, Dinamarca. **Proceeding...** Heidelberg: Springer, 2009. p.66-74.

GERALDO, A. P.; MOREIRA, V. P.; GONÇALVES, M. A. On-demand associative cross-language information retrieval. In: International Symposium on String Processing and Information Retrieval, 16., 2009, Saariselkä, Finland. **Proceeding…** Heidelberg: Springer-Verlag, 2009. p.165-173.

GEY, F. C.; KANDO, N.; PETERS, C. Cross-language information retrieval: the way ahead. **Inf. Process. Manage.**, v.41, n.3, p.415-431, 2005.

GEY, F.; JIANG, H. English-German cross-language retrieval for the GIRT collection-exploiting a multilingual thesaurus. Text REtrieval Conference, 8., 1999, Gaithersburg, Maryland. **Proceedings…** Gaithersburg, Maryland: NIST/DARPA, 1999. p.219-234.

GILLMEISTER, P. R. G.; CAZELLA, S. C. Uma análise comparativa de algoritmos de regras de associação: minerando dados da indústria automotiva. **Escola Regional de Banco de Dados**, Caxias do Sul, 2007. 10p.

GLOBAL REACH. **Global Internet statistics**. Disponível em: <http://global-reach.biz/globstats/index.php3>. Acesso em: 30 ago. 2007.

GOOGLE Basic of Search Grunnleggende om Google Disponível em: http://www.google.com/support/websearch/bin/answer.py?hl=no&answer=35889> acessado em 21 nov. 2007

GOOGLE-TRANSLATE. **Google-translate 2008**. Web search. Disponível em: <http://www.google.com/tanslate>. Acesso em: 21 nov. 2007.

GRAVANO, L. et al. Text joins in an RDBMS for web data integration. International Conference on World Wide Web, 12., 2003, Budapest, Hungary. **Proceedings…** New York, USA: ACM, 2003. p.90-101.

GREFENSTETTE, G. **Cross-language information retrieval**. Boston: Kluwer Academic Publishers, 1998. 200p.

HAYURANI, H.; SARI, S.; ADRIANI, M. Evaluating Language Resources for CLEF 2007. In: Advances in Multilingual and Multimodal Information Retrieval: Workshop of the Cross-Language Evaluation Forum, 8., 2007, Budapest, Hungary. **Revised Selected Papers**… New York, USA: Springer-Verlag, 2008.

HULL, D. Using statistical testing in the evaluation of retrieval experiments. In: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 16., 1993, Pittsburgh, Pennsylvania, United States. **Proceedings…** New York, USA: ACM, 1993.

KÜRSTEN, J.; WILHELM, T.; EIBL, M. The XTRIEVAL framework at CLEF 2007: domain-specific track. In: Advances in Multilingual and Multimodal Information Retrieval: Workshop of the Cross-Language Evaluation Forum, 8., 2007, Budapest, Hungary. **Revised Selected Papers**… New York, USA: Springer-Verlag, 2008. p.174-181.

LARSON, R. R. Logistic regression for metadata: Cheshire takes on adhoc-TEL. In: **Cross-Language Evaluation Forum**, 9., 2008, Aarhus, Dinamarca. **Proceeding...** Heidelberg: Springer, 2009.

LEC POWER TRANSLATOR. Desenvolvido por LEC Power Translator. Disponível em: <http://www.lec.com/default.asp>. Acesso em: jun. 2006.

LEMUR. Desenvolvido por Lemur. Disponível em: <http://www.lemurproject.org/>. Acesso em: 8 abr. 2009.

LIU, Y.; JIN, R.; CHAI, J. Y. A maximum coherence model for dictionary-based cross-language information retrieval. In: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 28., 2005, Salvador, Brazil. **Proceedings…** New York, USA: ACM, 2005. p.536-543.

LOVINS, J. B. Development of a stemming algorithm. T**ranslation and Computational Linguistics**, v.11, n.1, p.22-31, 1968.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge, Reino Unido: Cambridge, 2008. 547p.

MCNAMEE, P.; MAYFIELD, J. Scalable multilingual information access. In: **Cross-Language Evaluation Forum**, 2002, Roma, Italia. **Proceedings…** Heidelberg: Springer, 2003.

MOULINIER, I.; MOLINA-SALGADO, H. Thomson legal and regulatory experiments for CLEF 2002. In: **Cross-Language Evaluation Forum**, 2002, Roma, Italia. **Proceedings…** Heidelberg: Springer, 2003.

NGUYEN, D. et al. WikiTranslate: query translation for cross-lingual information retrieval using only Wikipedia. In: Cross-Language Evaluation Forum, 9., 2008, Aarhus, Dinamarca. **Proceeding...** Heidelberg: Springer, 2009.

OGAWA, Y.; MORITA, T.; KOBAYASHI, K. A fuzzy document retrieval system using the keyword connection matrix and a learning method. **Fuzzy Sets Syst.**, v.39, n.2, p.163-179, 1991.

ORENGO, V. M. **Assessing relevance using automaticall translated documents for cross-language information retrieval**. 2004. 258p. PhD (Dissertation) – School of Computer Science, Middlesex University, Londes, Reino Unido, 2004.

ORENGO, V. M.; BURIOL, L. S.; COELHO, A. R. A Study on the use of stemming for monolingual ad-hoc portuguese information retrieval. In: Workshop of the Cross-Language Evaluation Forum, 7., 2006, Alicante, Spain. **Revised Selected Papers…** Berlin: Springer, 2007. p.91-98.

ORENGO, V. M.; HUYCK, C. R. A stemming algorithm for Portuguese language. In: International Symposium on String Processing and Information Retrieval, 8., 2001, Laguna de San Raphael, Chile. **Proceedings…** Laguna de San Raphael, Chile: SPIRE, 2001. p.183-193.

ORENGO, V. M.; HUYCK, C. R. Portuguese-English cross-language information retrieval using latent semantic indexing. In: Cross-Language Evaluation Forum, 2002, Roma, Italia. **Proceedings…** Heidelberg: Springer, 2003.

PETERS, C. Working notes for the CLEF 2008 Workshop. In: Cross-Language Evaluation Forum, 9., 2008, Aarhus, Dinamarca. **Proceeding...** Heidelberg: Springer, 2009.

PORTER, M. F. Algorithm for suffix stripping. **Program**, v.14, n.3, p.130-137, 1980.

ROBERTSON, S. E. et al. Okapi at TREC-3. In: Text REtrieval Conference, 3., 1994, Gaithersburg, USA. **Proceedings…** Gaithersburg, USA: NIST/DARPA, 1994.

ROBERTSON, S. E.; SPARCK-JONES, K. Relevance Weighting of Search Terms. **Journal of the American Society for Information Science**, v.27, n.3, 1976.

ROSS, W.; PHILIP, H. Using the cosine measure in a neural network for document retrieval. Annual International ACM SIGIR conference on Research and development in information retrieval, 14., 1991, Chicago, United States. **Proceedings…** New York, USA: ACM, 1991. p. 202-210.

SALTON, G. Relevance feedback and the optimisation of retrieval effectiveness. In: _____. (Ed.). **The SMART retrieval system**: experiments in automatic document. Englewood Cliffs, Now Jersey: Prentice Hall, 1971. p. 324-336.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Communication ACM**, v.18, n.11, p.613-620, 1975.

SANDERSON, M. Word sense disambiguation and information retrieval. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 17., 1994, Dublin, Ireland. **Proceedings…** New York: ACM, 1994.

SAVOY, J. Data fusion for effective European monolingual information retrieval. In: Cross-Language Evaluation Forum, 2004, Bath, UK. **Proceedings…** Heidelberg: Springer, 2005. p.233-244.

SPARCK-JONES, K. What might be in a summary? **Information retrieval**, n. 93, p.9-26, 1993.

SPARCK-JONES, K.; RIJSBERGEN, C. J. V. **Report on the need for and provision of an 'ideal' test collection**. Cambridge: University Computer Laboratory, 1975.

SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL (SIGIR). Desenvolvido por SIGIR. Disponível em: <http://www.sigir.org/>. Acesso em: 15 jun. 2008.

THE TEXT RETRIEVAL CONFERENCE (TREC). Desenvolvido por TREC. Disponível em: <http://trec.nist.gov>. Acesso em: 11 out. 2007.

WEISS, S. M.; INDURKHYA, N. **Predictive data mining**. Boston: Morgan Kaufmann. 1998.

WIJSEN, J.; MEERSMAN, R. On the Complexity of Mining Quantitative Association Rules. **Data Mining and Knowledge Discover**, v.2, n.3, p.263-281, 1998.

WIKIPÉDIA. **Tf-idf**. Disponível em: <http://en.wikipedia.org/wiki/Tf-idf>. Acessado em: 10 mar. 2008.

S**tudent's t-test**. Disponível em: <http://en.wikipedia.org/w/index.php?title=Student%27s_t-test&oldid=352225215>. Acesso em: 5 mar. 2009.

I**nformation retrieval**. Disponível em : <http://en.wikipedia.org/w/index.php?title=Information_retrieval> Acesso em: 29 mar. 2010.

XU, J.; CROFT, W. B. Query expansion using local and global document analysis. Annual International ACM SIGIR Conference on Research and Development In Information Retrieval, 19., 1996, Zurich, Switzerland. **Proceedings…** New York, USA: ACM, 1996. p.4-11.

ZETTAIR. Desenvolvido por ZETTAIR. Disponível em: <http://www.seg.rmit.edu.au/zettair/>. Acesso em: 17 out. 2007.

ZHANG, Y.; VINES, P.; ZOBEL, J. Chinese OOV translation and post-translation query expansion in chinese--english cross-lingual information retrieval. **ACM Transactions on Asian Language Information Processing (TALIP)**, v.4, n.2, p.57-77, 2005.

ZHOU, D. et al. NTCIR-6 Experiments using pattern matched translation extraction. In: NTCIR Workshop Meeting, 6., 2007, Tokio, Japão. **Proceeding…** Tokio, Japão: NILL, 2007.

ZHOU, D.; TRURAN, M.; BRAILSFORD, T. Ambiguity and unknown term translation in CLIR. In: Advances in Multilingual and Multimodal Information Retrieval: Workshop of the Cross-Language Evaluation Forum, 8., 2007, Budapest, Hungary. **Revised Selected Papers**… New York, USA: Springer-Verlag, 2008.