UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Um modelo difuso de recuperação de documentos utilizando processamento morfológico

Bernd Heinrich Storb

Dissertação apresentada para obtenção do grau de Mestre em Engenharia da Produção

Orientador
Raul S. Wazlawick

setembro/1997

Um modelo difuso de recuperação de documentos utilizando processamento morfológico

Bernd Heinrich Storb

Essa dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia, especialidade em Engenharia de Produção e aprovada em sua forma final pelo curso de Pós-Graduação em Engenharia de Produção.

Prof. Ricardo M. Barcia, PhD
(Coordenador do Curso)

Banca Examinadora:

Prof. Luiz Fernando J. Maia, Dr.
(Presidente)

Prof. Ruth E. V. Lopes, MSc

Prof. Rogério Cid Bastos, Dr.

Prof. Roberto C. d. S. Pacheco, Dr.

DEDICATÓRIA

Este trabalho é dedicado a minha esposa Mareni Rocha Farias, por seu amor, sua compreensão, companheirismo, carinho e a paciência sem limite.

AGRADECIMENTOS

Ao professor Raul Wazlawick, por sua orientação e paciência.

Ao professor Antônio Carlos Mariani, pelo auxílio e sugestões durante o desenvolvimento do trabalho.

À professora Ruth Elisabeth Lopes, por sua atenção e esclarecimentos das dúvidas na área de lingüística.

Ao amigo Adrian Ferreira pela amizade e apoio via Internet.

À minha esposa Mareni, por sua paciência na leitura deste trabalho.

A todos, que de alguma forma, colaboraram para a finalização de mais esta etapa.

RESUMO

O crescimento de armazenamento de informações em meios digitais torna também a área de recuperação automática de informação mais importante. Documentos representam um tipo de informação, cuja recuperação apresenta problemas específicos. A recuperação de documentos pode ser definida como a seleção de documentos, caracterizados por um conjunto de descritores (palavras-chave ou outros símbolos), como resposta a uma consulta. A recuperação difusa de documentos refere-se a métodos baseados na teoria de conjuntos difusos.

Esta dissertação propõe um modelo de recuperação de documentos para a língua portuguesa desenvolvido a partir de um modelo de Miyamoto de recuperação difusa de documentos. O modelo de Miyamoto baseia-se na detecção de similaridades semânticas entre descritores através de co-ocorrências. Esta proposta pode ser considerada uma extensão do modelo de Miyamoto, considerando também similaridades lexicais. Nesta extensão, descritores e consultas são consideradas expressões, i.e. seqüências de palavras e conectores. A similaridade entre palavras é determinada através de comparações entre possíveis radicais, determinados por um sistema de reconhecimento de radicais e sufixos, para detectar palavras com significado idêntico ou semelhante . A detecção de similaridades entre expressões é realizada por meio de uma adaptação de um modelo de Bruza e van der Weide comparando as palavras e os conectores. A extensão proposta para o modelo de Miyamoto considera, por um lado, a determinação de um thesaurus difuso através de um índice difuso determinado pela similaridade dos descritores dos documentos e, por outro lado, a possibilidade de utilização de um vocabulário não-controlado através da determinação de similaridades entre descritores dos documentos e expressões de consultas.

Para a avaliação do modelo proposto, foi feita uma implementação, bem como do modelo de Miyamoto, em Smalltalk. Estabeleceu-se uma pequena base de documentos para a realização de testes, comparando-se os dois modelos.

Os resultados dos testes indicam que a extensão do thesaurus difuso leva a uma maior qualidade na recuperação dos documentos e que a possibilidade de utilização de um vocabulário não-controlado leva a uma perda na qualidade dos resultados da recuperação, mantendo, no entanto, a precisão nos níveis de evocação.

Estes resultados demonstram a possibilidade de utilização do modelo proposto na construção de um thesaurus difuso para a língua portuguesa. Quanto á possibilidade de utilização de um vocabulário não-controlado, o modelo mostra-se viável para pequenas bases de documentos, podendo ser ainda aperfeiçoado, no que diz respeito à determinação das similaridades entre expressões.

ABSTRACT

The increase of digital information storage turn the automatic information retrieval still more important. Documents are a kind of information with specific retrieval problems. Document retrieval may be defined as the selection of documents, characterized by a set of descriptors (keywords or other symbols), matching with a query. The fuzzy document retrieval refers to methods that are based on the theory of fuzzy sets.

This dissertation reports a model of document retrieval for Portuguese language, developed from a document retrieval model of Miyamoto. The Miyamoto model is based upon semantic similarity detection of descriptors by co-occurences. The proposed model may be considered an extension of the Miyamoto model by taking into account also lexical similarities. Hence, descriptors and queries are expressions, i.e. series of words and connectors. The similaritiy between words is based on the comparison between possible radicals, determined by a recognition system of radicals and suffixes, for detection of words with identical or similar meanings. The expression similarity detection is done by an adaptation of a Bruza and van der Weide model, comparing words and connectors. The proposed Miyamoto model extension consider both the determination of a fuzzy thesaurus by a fuzzy index determinated through lexical descriptor similarities and the possibility of a non-controlled vocabulary use by determination of similarities between document descriptors and query expressions.

The evaluation of the model proposed here and the Miyamoto one was done by implementing them in Smalltalk. A small document base was created for the comparison of the models.

The tests results indicate a higher quality in document retrieval for the fuzzy thesaurus extension. The use of a non-controlled vocabulary reduced the quality of retrieval, but maintained the precision in recall levels.

The obtained results show the usefulness of the proposed model to fuzzy thesaurus construction for the Portuguese language. Furthermore, the results indicate viability for the non-controlled vocabulary use to small document bases. However the model may be improved in expression similarity determination.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE EXEMPLOS

APRESENTAÇÃO

Este trabalho visa a apresentar uma proposta de um modelo de recuperação automática de documentos utilizando duas ferramentas da inteligência artificial: elementos de processamento de linguagem natural e teoria de conjuntos difusos.

Inicialmente serão abordados alguns aspectos teóricos relacionados a estas duas ferramentas. No que diz respeito à linguagem natural, serão enfocadas especialmente as questões relacionadas à análise morfológica. Quanto à teoria de conjuntos difusos, serão discutidas as operações padrão sobre conjuntos difusos, a composição de relações difusas e as operações lógicas e/ou.

Posteriormente serão apresentados alguns sistemas desenvolvidos para a recuperação automática de documentos. Dentre estes sistemas, será aprofundado o sistema de Miyamoto, o qual foi utilizado no sistema proposto neste trabalho.

Na seqüência será apresentado o modelo desenvolvido o qual emprega elementos de um analisador morfológico que determina radicais e sufixos de palavras desconhecidas, que é utilizado na análise de expressões. O resultado desta análise é utilizado por um sistema de recuperação difusa semelhante ao sistema descrito por Miyamoto.

Por fim, são apresentados métodos de avaliação de um sistema de recuperação de documentos, bem como uma comparação entre o modelo de Miyamoto e o modelo proposto para uma pequena coleção de documentos.

1. INTRODUÇÃO

Este capítulo visa a apresentar aspectos relacionados ao problema da recuperação de documentos, a justificativa e a importância de trabalhos nesta área, bem como o objetivo da presente proposta.

1.1. Identificação do problema

A explosão recente do provimento de informação na Internet, e o crescimento de busca de informações por usuários, trouxe certamente novos desafios para a área da recuperação de informação [SPA 96]. A quantidade de informações armazenadas em base de dados cresce em velocidade acelerada e, atualmente, o tamanho das bases de dados é medido em gigabytes e terabytes [HED 95]. Além disso a rede é um lugar onde todo mundo é gerador de informação, como escreveu James Gleick, o criador da Teoria do Caos, em um artigo para The New York Times.

A verdadeira identidade da Internet é um conjunto de software de protocolos ou aplicativos, milhares de servidores de dados e mais do que 20 milhões de usuários procurando os dados necessitados, no entanto, muitos confundem esta estrutura com a estrutura física da rede [REI 94]. Para muitos usuários da Internet, "uma vez ligado na rede está tudo resolvido e é só surfar...". Bem, não é tão fácil assim, segundo Penteado, em um artigo na revista Informática Exame [PEN 95], "...gigantesca e caótica, a Internet é um ótimo lugar para ficar perdido, principalmente na sua região gráfica, a Web. Para que isso não aconteça, existem serviços que ajudam o interessado a encontrar seus assuntos preferidos ou necessários. São como uma bússola,..., a orientar os usuários entre as mais de 50000 redes espalhadas pelo mundo todo...". Com isso a autora refere-se a serviços como o Yahoo, Webcrawler ou Alta Vista.

Mas não somente na Internet o gerenciamento de informações ou mais especificamente a recuperação de documentos representa um grande desafio, diversas áreas necessitam de sistemas deste tipo em sua rotina de trabalho, como mostra o exemplo citado na revista Informática Exame [NAS 95]: "Todos os dias, funcionários da importadora Polimate, de Porto Alegre, atendem de dez a vinte telefonemas de clientes. Eles querem saber, por exemplo, se as mercadorias que compraram já embarcaram no porto de origem e, principalmente, quando devem chegar ao Brasil. Outras vezes, os clientes precisam da especificação exata de algum artigo importado um ano atrás. E nenhum deles, é claro, gosta de ficar esperando ao telefone. As respostas a tais perguntas poderiam ser rápidas caso a burocracia dos processos de importação não criasse tantos documentos, espalhando informações por todos eles. Cada processo é

composto, em média, por cinqüenta documentos, entre eles as guias emitidas pelo Banco do Brasil. Com um total de 500 processos de importação por ano, dos quais de quarenta a cinqüenta sempre em andamento, os setenta funcionários da Polimate naufragavam em pilhas de papéis quando precisavam fornecer informações. Às vezes, demoravam dias para localizá-las, lendo e relendo guias, ofícios, carimbos e autenticações em busca da informação certa." Segundo a revista o problema foi resolvido através de um sistema digitalizado que garante a resposta ao cliente geralmente na hora.

Segundo o mesmo artigo "um executivo gasta perto de três horas por semana apenas procurando documentos. Uma secretária gasta quatro horas semanais nessas mesmas tarefas. Todos os dias, são criados e arquivados cerca de 200 milhões de documentos só nos EUA". Estes dados demonstram que a recuperação de documentos é um grande problema a ser solucionado por qualquer empresa preocupada em elevar sua produtividade [NAS 95].

Outra aplicação da recuperação automatizada de informações é na própria área da informática. Uma das vantagens sempre apontada, tanto da programação estruturada, quanto da programação orientada a objetos é a possibilidade de reuso de componentes uma vez implementados e testados. Mas, segundo Frakes & Gandel [FRA 89], um dos problemas fundamentais em relação a reusabilidade de software é a falta de ferramentas para representar, indexar, armazenar e recuperar componentes reusáveis . Por exemplo, uma biblioteca enorme de rotinas tem pouca utilidade, se não existe uma ferramenta que permita encontrar um módulo que resolva um problema específico.

Contudo, um dos maiores problemas, ainda não solucionado, na recuperação de documentos, apontado por Croft [CRO 83], é a interface dos usuários. Sistemas utilizando "strings" e operadores booleanos colocam uma carga pesada em cima dos usuários durante o processo de formulação das consultas. Experiências com sistemas bibliográficos mostram que usuários tendem ou a formular consultas muito simples ou a delegar a pesquisa a um intermediário treinado. Sendo assim, existe uma demanda crescente por trabalhos na área de recuperação de informação, não só do ponto de vista de aplicativos, como também no desenvolvimento de novos modelos.

1.3. Objetivo do trabalho

O objetivo deste trabalho é a apresentação e avaliação de um modelo de recuperação de documentos, que por um lado utiliza a teoria difusa e por outro lado recursos de processamento de linguagem natural. Este modelo é uma extensão de um modelo difuso de recuperação de documentos de Miyamoto [MIY 90], que utiliza um sistema de reconhecimento de radicais e sufixos [STO 96]. Ao contrário do modelo de Miyamoto, o modelo proposto permite consultas sem utilização de um vocabulário controlado.

1.3. Justificativa e importância do tema

Segundo Sparck-Jones & Galliers [SPA 96], considerando um sistema de recuperação bibliotecária, parece ser óbvio que o objetivo é retornar ao usuário o que ele necessita, o que por conveniência é chamado "documentos relevantes". Mas o que é óbvio à primeira vista não é obvio numa inspeção mais profunda:

- muitas vezes os usuários procuram informações em áreas que eles ainda não conhecem;
- geralmente um serviço de informação é procurado por vários tipos de usuários com diferentes necessidades.

Um dos problemas fundamentais em recuperação de documentos é que as necessidades dos usuários, isto é, o que faz um documento relevante para uma consulta não é um fato objetivo, completamente acessível.

Segundo Klir e Yuan [KLI 95], o uso da teoria de conjuntos difusos na recuperação de informação tem pelo menos as duas vantagens em comparação com os métodos clássicos:

- relações de relevância difusas e thesauri difusos são mais expressivas;
- a construção das relações de relevância e dos thesauri é mais realística.

Além disso, segundo Cox em [BAR 93], é fácil projetar, criar e validar conjuntos difusos, e por diferentes razões, como a tolerância a falhas, eles são extremamente robustos.

Segundo Jarke *et. al.* [JAR 85] não se pode mostrar uma superioridade na precisão de consultas e na efetuação de tarefas de sistemas com interfaces em linguagem natural sobre sistemas com interfaces em linguagem formal. Mas por outro lado, segundo Frakes & Gandel [FRA 89], experimentos mostraram que sistemas de recuperação de documentos sem vocabulário controlado produzem resultados comparáveis com os de sistemas com vocabulário controlado. Além disso, segundo Jarke *et al.* [JAR 85], sistemas com interfaces em linguagem natural requerem um tempo menor na formulação de consultas e não necessitam um treinamento dos usuários.

Em sistemas que não trabalham em áreas específicas como por exemplo Yahoo, Webcrawler e Alta Vista na Internet, também não é realístico supor que os usuários poderiam conhecer o vocabulário controlado. Assim, consultas baseiam-se muitas vezes na estratégia de tentativa e erro.

2. FUNDAMENTOS TEÓRICOS

Neste capítulo serão abordados fundamentos teóricos relacionados ao *Processamento de Linguagem Natural* e à *Teoria Difusa*, os quais foram empregados no desenvolvimento do modelo proposto. Serão apresentados apenas os conceitos utilizados, sem a pretensão de esgotar o assunto.

2.1. Processamento de Linguagem Natural

Entre os diversos campos de aplicação aos quais a Inteligência Artificial empresta suas técnicas, o Processamento da Linguagem Natural figura como um dos maiores desafios desta disciplina. O objetivo maior do Processamento da Linguagem Natural é tratar a língua de maneira automática através de formalismos que explicitem os conhecimentos lingüísticos, tornando-os operacionais e calculáveis ou, em outros termos, passíveis de serem tratados por computador.

Nos últimos anos, a informação tornou-se uma fonte de recursos essencial, adquirindo a mesma importância dos recursos materiais, financeiros e humanos [FAV 95]. A exemplo da chamada globalização da economia, a dinâmica da interação entre os povos tem demonstrado, cada vez mais, a necessidade de computadores capazes de tratar as línguas naturais, permitindo classificar, analisar, resumir, interrogar e traduzir documentos já existentes e, também, compor e estruturar novos documentos. Além disso, a informática está cada vez mais presente no cotidiano das pessoas comuns, não especialistas. Sendo assim, a exigência por mecanismos que permitam ao usuário comum dialogar com um computador usando sua própria linguagem, tem sido cada vez maior.

As línguas são os veículos das informações científicas, tecnológicas e industriais. No futuro, somente as línguas dispondo de sistemas eficientes de tratamento informatizado terão acesso a estas informações e delas serão os veículos privilegiados [FAV 95].

2.1.1. Modelos lingüísticos

Lingüística é a disciplina que estuda a linguagem e as línguas particulares. A língua pode ser considerada como um sistema de valores e um conjunto de convenções necessárias, adotadas por uma comunidade [SIL 89]. Ela é caracterizada por um sistema de signos vocais distintos e significativos. A língua está depositada como produto social na mente de cada falante e é, por excelência, o veículo do conhecimento humano e a base do patrimônio cultural de um povo. A fala é a forma pela qual o indivíduo utiliza a língua.

No processamento da linguagem natural em informática faz-se necessária, consciente ou inconscientemente, a formulação de hipóteses sobre o que é linguagem, as quais se traduzem na escolha de um modelo. Qualquer modelo de representação de conhecimento é a *priori*, pertinente. Segundo Coulon & Kayser [COU 92], os modelos que têm sido efetivamente utilizados no processamento de dados em linguagem natural são:

- **modelos que não se utilizam da sintaxe:** baseados em hipóteses restritivas, mas simples, considerando as palavras em duas categorias: as portadoras de significação e as demais.
- **modelos que se utilizam da sintaxe:** caracterizam-se pelo fato de reconhecerem que as sentenças têm uma estrutura, a qual desempenha um papel importante para a compreensão de textos.
- **modelos lógicos:** propõem-se a representar o significado dos textos sob uma forma que convém a processamentos informáticos eficazes. Podem ser gerados ao longo de uma análise sintática conduzida a partir de um dos modelos anteriores.
- **modelos psicocognitivos**: consideram as intuições e, portanto, não negligenciam as deduções. Formalizar isto requer, por sua vez, um grande número de regras de inferências ligadas às diferentes situações da vida real, sabendo-se que o resultado destas inferências não constitui uma verdade lógica, mas apenas uma conseqüência esperada.

Nos processamentos executados para interpretar um enunciado, geralmente são encontrados módulos de serviço para as análises morfológica e lexical, que evitam que se tenha de mencionar num léxico todas as formas possíveis que uma palavra possa assumir.

2.1.2. Morfologia lingüística

A conceituação de morfologia está na dependência direta do modelo adotado pelo cientista. Seu objeto de estudo, segundo Anderson [AND 91], é a estrutura da palavra e os caminhos, através dos quais, sua estrutura reflete sua relação com outras palavras. Dentro da visão estruturalista as palavras são compostas por unidades mínimas dotadas de significado, denominadas morfemas. Os morfemas são os elementos centrais do código lingüístico e sua análise é um dos problemas fundamentais da lingüística teórica. Segundo Cabral, as definições de morfologia apresentadas por alguns autores poderiam ser reunidas em: *parte da gramática que descreve as unidades mínimas de significado, sua distribuição, variantes e classificação, conforme as estruturas onde ocorrem, a ordem que ocupam, os processos na formação de palavras e suas classes* [CAB 85].

A análise mórfica consiste na descrição do vocábulo mórfico, depreendendo suas formas mínimas ou morfemas, de acordo com uma significação e uma função elementar que lhes são atribuídas dentro da significação e da função total do vocábulo [SIL 89].

2.1.2.1. Classificação das palavras

As palavras podem ser distribuídas em dez grupos, de acordo com os critérios nocionais, semânticas e estruturais que indicam:

- **substantivo**: toda a palavra que especifica substância, ou seja, coisa que possua existência, quer animada, inanimada, real, imaginária, concreta ou abstrata;
- **artigo**: palavra que tem por fim individualizar a coisa: *a, o, um, uma*;
- **adjetivo**: palavras que expressam a qualidade ou características dos seres;
- **numeral**: encerram a idéia de número: *um, dois, ambos, sexto*;
- **pronome**: palavras que substituem ou podem substituir um nome, um substantivo: *ele, que, quem*;
- **verbo**: palavra que exprime ação, estado, fato ou fenômeno;
- **advérbio**: palavra que modifica o sentido do verbo, do adjetivo e do próprio advérbio;
- **preposição**: palavras que ligam um termo dependente a um termo principal, estabelecendo uma relação entre eles;
- **conjunção**: palavras que servem para ligar, não palavras, como a preposição, mas orações;
- **interjeição**: exprimem manifestações súbitas, repentinas, momentâneas do nosso íntimo: *Ai! Oh!*

2.1.2.2. Flexão das palavras

No que diz respeito à flexão, ou seja, à propriedade que certas classes de palavras têm de sofrer alteração, em português, na parte final, podemos dividir as dez classes de palavras em dois grandes grupos:

- **invariáveis**: palavra que não flexiona, isto é, não sofre nenhuma alteração na última sílaba:
    - advérbio
    - preposição
    - conjunção
    - interjeição
- **variáveis**: palavras que sofrem alteração na última sílaba:
    - substantivo
    - artigo
    - adjetivo
    - numeral
    - pronome
    - verbo

2.1.2.3. A estrutura das palavras

Do ponto de vista da estrutura das palavras, podemos considerar que estas são formadas por:

- **radical, tema**: elementos básicos significativos da palavra, despojada de seus elementos secundários (quando houver):
  - **radical**: elemento básico e significativo das palavras, consideradas sob o aspecto gramatical, semântico e pragmático, dentro da língua portuguesa atual (palavra despojada de seus elementos secundários)

CERT-eza, in-CERT-eza, CAFE-teira, RECEB-er, ex-PORT-ação;

  - **tema**: radical acrescido de um constituinte temático

CANTA-r, BATE-r, DEVE-dor, FINGI-mento, PERDOÁ-vel;

- **afixos (sufixo e prefixo), desinência, constituinte temático**: elementos modificadores da significação e/ou da classificação do radical:
  - **prefixo:** elemento anteposto ao radical
  - **sufixo**: elemento posposto ao radical

IN (prefixo) - AT (radical) - IVO (sufixo)
EM (prefixo) - POBR (radical) - ECER (sufixo);

  - **desinência ou sufixo flexional**: elemento terminal indicativo de flexão
    - *nominais*: indicam as flexões de gênero (masculino, feminino)

menin-O, menin-A;

    - *de número*: indicam as flexões de número (singular e plural)

menino-S, menina-S

    - verbais: indicam as flexões de número, pessoa, modo e tempos verbais

ama-S, ama-MOS, ama-IS, ama-M, ama-VA, ama-VAS, etc.

  - **constituinte temático**: elemento que, acrescido ao radical, forma o tema de nomes e verbos

and-A-r, bat-E-r, part-I-r, cafe-T-eira;

- **vogal de ligação**: vogais de ligação que servem como conectores entre radicais.
  agr-I-cultor, fot-O-grama.

Assim, segundo Villalva [VIL 94], não considerando prefixação e composição de radicais, a estrutura composicional das palavras pode ser representada como na figura 2.1.

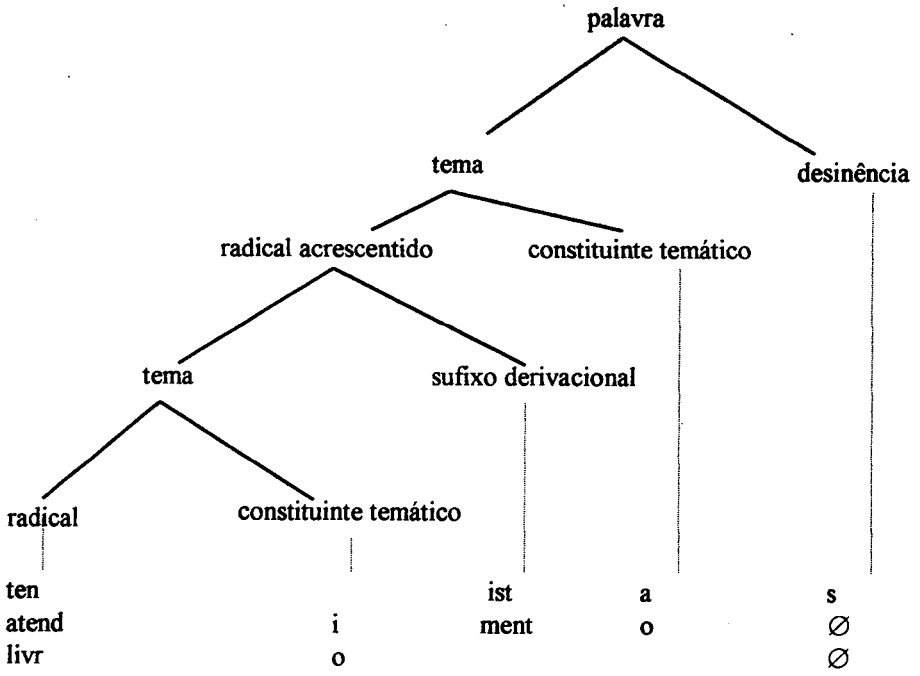


Figura 2.1: Representação das estruturas morfológicas das palavras 'tenistas', 'atendimento' e 'livro'.

2.1.2.4. A formação das palavras

Quanto à formação, as palavras podem ser consideradas primitivas ou derivadas, ou ainda, segundo o radical, em simples ou compostas.

- **palavras primitivas**: não derivam de outras dentro da língua portuguesa,
  pedra, terra, dente, pobre, etc.

- **palavras derivadas**: provêm de outras por derivação, podendo ser realizadas a partir de quatro maneiras diferentes:

  - **por prefixação**: antepondo-se um prefixo a um radical,

IN-capaz, SUPER-sônico, DES-ligar;

  - **por sufixação**: acrescentando um sufixo ao radical,

dent-ISTA, joga-DOR, feliz-MENTE;

  - **por derivação parassintética ou parasíntese**: anexando-se ao mesmo tempo um prefixo e um sufixo a um nome. Os vocábulos parassintéticos são quase sempre verbos e têm como base um substantivo ou um adjetivo

A-list-AR, EM-vergonh-AR, DES-alm-ADO, A-maci-AR, ES-fri-AR;

  - **por derivação regressiva**: substituindo-se a terminação de um verbo pelas desinências -A, -O ou -E,

mudar→mudA, chorar→chorO, atacar→ataquE;

- **palavras simples**: são as que têm só um radical,
  beleza, livre, recomeçar, etc.

- **palavras compostas**: são as que apresentam mais de um radical. A composição de uma palavra pode ser por justaposição ou aglutinação:

  - **por justaposição**: unindo-se duas ou mais palavras, sem lhes alterar a estrutura

passatempo, vaivém, girassol, greco-latino, etc.

  - **por aglutinação**: unindo-se duas ou mais palavras, com perda de um ou mais fonemas

aguardente (água+ardente), embora (em+boa+hora), planalto (plano+alto), etc.

- **redução**: algumas palavras apresentam ao lado de sua forma plena, uma forma reduzida
  auto (automóvel), cinema (cinematografia), foto (fotografia), etc.

2.1.2.5 O significado das palavras

Segundo Ilari & Geraldi [ILA 85], o significado de uma palavra é o conjunto de contextos lingüísticos em que pode ocorrer. Conceitos que descrevem a semelhança de significados semânticos de palavras são:

- **sinonímia**: duas palavras são sinônimas sempre que podem ser substituidas no contexto de qualquer frase sem que a frase passe de falsa a verdadeira, ou vice-versa, por exemplo calvo e careca. Mas segundo [ILA 85], a sinonímia é um fenômeno gradual nas linguagens naturais.
- **hiponímia**: a relação hiponímica é aquela que intercorre entre expressões com sentido mais específico e expressões genéricas; por exemplo, a relação entre pardal e passarinho: todo pardal é um passarinho, mas nem todo passarinho é um pardal.
- **antonímia**: a relação entre palavras que representam sentidos incompatíveis com a mesma situação, por exemplo, branco/preto, colorido/incolor, bom/mau, chegar/partir, abrir/fechar, nascer/morrer, todo/nenhum.
- **ambigüidade e polissemia**: duplicidade de sentido de uma única palavra, por exemplo, banco pode ser uma casa bancária ou um assento de jardim.

Segundo Said Ali em [VIL 94], "a derivação toma palavras existentes e lhes acrescenta certos elementos formativos com que adquirem sentido novo, referido contudo ao significado da palavra primitiva". Mas por outro lado, Villalva [VIL 94] afirma que, "note-se que em Português é possível detectar a coexistência de palavras de significação muito próxima, mas cuja estrutura morfológica é diferente: audaz e audacioso são formas em que ocorrem os sufixos -az e -oso".

2.1.3. Classe aberta e classe fechada

Em processamento de linguagem natural é comum dividir as categorias de palavras em duas classes: classe aberta e classe fechada:

- **classe aberta**: compreendem as principais categorias lexicais como verbos, substantivos e adjetivos [OUH 91], apresentando, entre outras, características:
  - no dicionário de uma língua, representam o maior número de palavras;
  - apresentam um papel temático criado pelos falantes;
- **classe fechada**: compreendem as demais categorias, também designadas categorias funcionais como artigos, numerais, pronomes, advérbios, preposições, conjunções e interjeições [OUH 91], apresentando, entre outras, características:
  - no dicionário de uma língua, representam um pequeno número de palavras;
  - não apresentam um papel temático.

2.1.4. Considerações finais sobre processamento de linguagem natural

O modelo de recuperação de documentos apresentado neste trabalho não considera a sintaxe de expressões de consulta. As palavras da classe aberta são consideradas como portadoras do significado, enquanto as palavras de classe fechada servem como conectores entre as palavras da classe aberta. Como estrutura das palavras da classe aberta considera-se, como em [STO 96], uma simplificação que

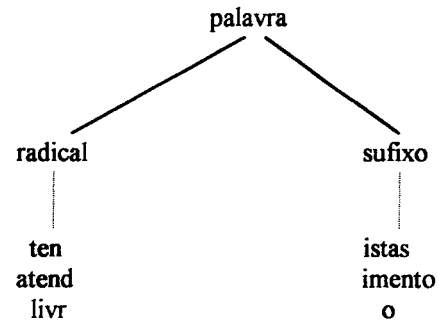


Figura 2.2: Divisão das palavras 'tenistas', 'atendimento' e 'livro' em radical e sufixo.

encara as palavras compostas de radicais e sufixos (figura 2.2). O radical pode ser composto de vários radicais e o sufixo pode ser uma composição de sufixos derivacionais e flexionais, constituintes temáticos e desinências.

Além disso considera-se que palavras com o mesmo radical possuem um significado semelhante. Isto quer dizer que, entre os vários significados de duas palavras, existem dois significados semelhantes, um para cada palavra. Por outro lado, considerando ambigüidades, sinomínia e hiponímia, pode ser que palavras com o mesmo radical possuam significados bem diferentes em um determinado contexto, e palavras com significados semelhantes podem ter radicais bem diferentes.

2.2. Teoria de incerteza difusa

Entre os vários paradigmas que surgiram na ciência neste século, encontra-se o conceito da incerteza. Do ponto de vista tradicional, a incerteza (imprecisão, não-especificação, vagueza, inconsistência, etc) era considerada não-científica. De acordo com a visão moderna, incerteza é considerada essencial para a ciência.

Existe um acordo geral que um ponto importante na evolução do conceito moderno da incerteza foi o artigo de Lofti A. Zadeh [ZAD 65]. Nesta publicação Zadeh introduziu uma teoria cujos objetos são conjuntos com limites imprecisos - os conjuntos difusos. A pertinência em um conjunto difuso não é uma questão de ser ou não-ser, mas, mais propriamente, de um grau de intensidade.

2.2.1. Funções de pertinência e conjuntos difusos

Conjuntos da teoria clássica podem ser identificados com suas funções características. A cada elemento do universo preestabelecido é atribuído o valor *1*, se o elemento é membro do conjunto considerado. Se o elemento não é membro do conjunto considerado é atribuido o valor *0*. Essas funções características podem ser generalizadas permitindo que elementos do universo sejam mapeados para valores do intervalo *[0,1]*, indicando o grau de pertinência de tal maneira que valores maiores denotam um grau de pertinência maior. Tais funções são chamadas funções de pertinência, e os conjuntos definidos assim, são chamados conjuntos difusos.

Duas notações diferentes para as funções de pertinência de um conjunto difuso A em um universo X são usualmente empregadas na literatura:

$$\mu_A : X \rightarrow [0,1] \text{ ; e } A : X \rightarrow [0,1] .$$

A primeira anotação foi utilizada por Zadeh em [ZAD 65], diferenciando entre a função de pertinência μ_A e o conjunto difuso A. A segunda anotação não faz esta distinção, utilizando o fato de que cada conjunto difuso é definido única- e completamente por uma única função de pertinência.

A figura 2.3 mostra um exemplo de dois conjuntos difusos A e B em um universo $X = [0,3]$.

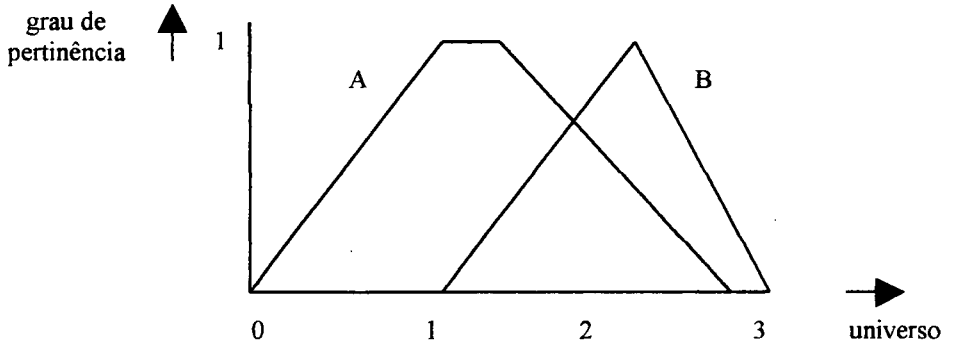


Figura 2.3: Exemplo de dois conjuntos difusos A e B definidos em um universo $X = [0,3]$.

Para conjuntos difusos definidos em um universo finito existe uma notação especial muito utilizada na literatura: Seja A um conjunto difuso no universo finito X, sejam x_i, $1 \leq i \leq n$, os elementos de X com grau de pertinência em A maior do que 0 e a_i denota o grau de pertinência de x_i em A. Então pode-se empregar a seguinte notação para o conjunto A:

$$A = a_1/x_1 + a_2/x_2 + \ldots + a_n/x_n$$

Além disso utiliza-se uma notação vetorial para conjuntos difusos. O mesmo conjunto difuso A pode ser anotado em forma vetorial como

$$\underline{x_1 \quad x_2 \quad . \quad . \quad . \quad x_n}$$

$$A = (a_1 \quad a_2 \quad . \quad . \quad . \quad a_n)$$

Para um conjunto difuso A definido num universo X define-se o suporte de A em X, *suporte*(A), como o conjunto de elementos de X, cujo grau de pertinência em A é maior do que zero, i.e.

$$suporte(A) = \{x \in X \mid A(x) > 0\}$$

2.2.1.1. Operações padrão sobre conjuntos difusos

As operações básicas da teoria dos conjuntos - complemento, intersecção e união - podem ser generalizadas para conjuntos difusos de diferentes formas (veja, por exemplo [KLI 95] cap. 3). No entanto, existem generalizações especiais, já introduzidas por Zadeh, normalmente chamadas operações padrão [ZAD 65].

O complemento padrão, A^C, de um conjunto difuso A em um universo X é definido por:

$$A^C(x) := 1 - A(x), \text{ para todos os } x \in X.$$

A figura 2.4 mostra um conjunto difuso B e seu complemento.

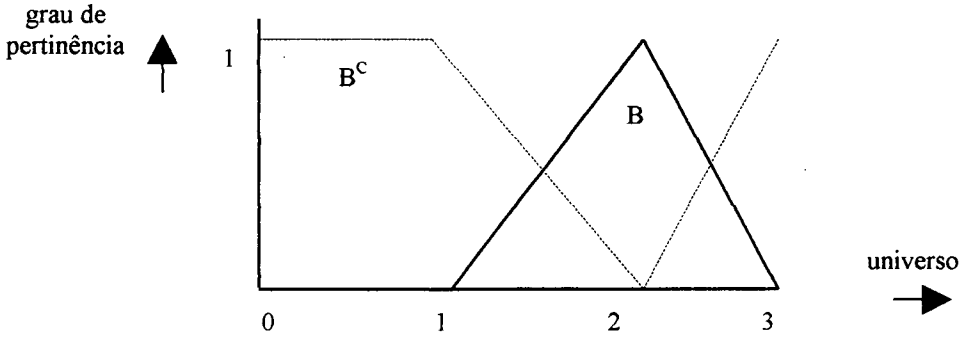


Figura 2.4: Um conjunto difuso B e seu complemento.

Sejam A e B dois conjuntos difusos em um universo X; a intersecção padrão, $A \cap B$, e a união padrão, $A \cup B$, são definidos para todos os $x \in X$, pelas equações:

$$(A \cap B)(x) := min[A(x);B(x)],$$

$$(A \cup B)(x) := max[A(x);B(x)].$$

A Figura 2.5 representa a união e intersecção padrão dos conjuntos difusos *A* e *B* da Figura 2.3.

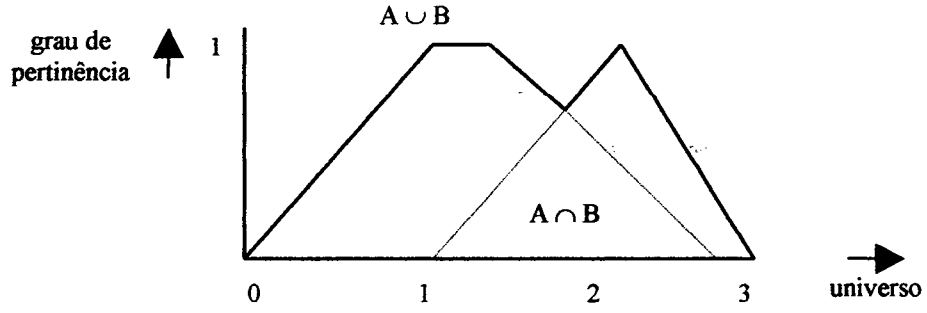


Figura 2.5: União e intersecção padrão dos conjuntos difusos *A* e *B* da figura 2.3.

Segundo Klir & Yuan [KLI 95], as operações padrão apresentam as seguintes propriedades fundamentais:

Involução	$(A^C)^C = A$
Comutatividade	$A \cup B = B \cup A$; $A \cap B = B \cap A$
Associatividade	$(A \cup B) \cup C = A \cup (B \cup C)$
	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributividade	$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
	$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
Idempotência	$A \cup A = A$; $A \cap A = A$
Absorção	$A \cup (A \cap B) = A$; $A \cup X = X$
	$A \cap (A \cup B) = A$; $A \cap \varnothing = \varnothing$
Identidade	$A \cap X = A$; $A \cup \varnothing = A$
leis de De Morgan	$(A \cap B)^C = A^C \cup B^C$
	$(A \cup B)^C = A^C \cap B^C$

onde *A*, *B* e *C* são conjuntos difusos no universo *X*.

Ao contrário da teoria de conjuntos clássica, as operações padrão difusas não satisfazem a lei da contradição e a lei do meio excluído, ou seja em geral

$$A \cap A^C \neq \varnothing \text{ e } A \cup A^C \neq X.$$

Além disso pode-se considerar uma inclusão difusa, que define uma ordem parcial na classe dos conjuntos difusos em um universo *X*. Sejam *A* e *B* dois conjuntos difusos em um universo *X*; *A* é chamado subconjunto de *B*, $A \subseteq B$, se e somente se $A(x) \leq B(x)$ para todos os $x \in X$.

2.2.1.2. Corte-α

Um dos conceitos mais importantes da teoria de conjuntos difusos é o conceito de corte-α. Dado um conjunto difuso A em um universo X e um número $\alpha \in [0,1]$; o corte-α, ${}^{\alpha}A$, é um conjunto não difuso que contém todos os elementos do universo X com grau de pertinência em A maior ou igual ao valor α, isto é:

$$ {}^{\alpha}A = \{ x \mid A(x) \geq \alpha \}. $$

Sejam A e B conjuntos difusos; α e β elementos do intervalo $[0,1]$. Os cortes-α satisfazem as seguintes propriedades fundamentais [KLI 95]:

$\alpha \leq \beta$ *implica em* ${}^{\alpha}A \subseteq {}^{\beta}A$
${}^{\alpha}(A \cap B) = {}^{\alpha}A \cap {}^{\alpha}B$
${}^{\alpha}(A \cup B) = {}^{\alpha}A \cup {}^{\alpha}B$
em geral ${}^{\alpha}(A^C) \neq ({}^{\alpha}A)^C$
$A \subseteq B$ *se e somente se* ${}^{\alpha}A \subseteq {}^{\alpha}B$ *para todos os* $\alpha \in [0,1]$
$A = B$ *se e somente se* ${}^{\alpha}A = {}^{\alpha}B$ *para todos os* $\alpha \in [0,1]$

O principal papel dos cortes-α na teoria dos conjuntos difusos consiste na capacidade de representar os conjuntos difusos. Eles fornecem uma conexão entre a teoria dos conjuntos difusos e a teoria clássica de conjuntos. Esta conexão é dada pelo teorema de decomposição [KLI 95], onde cada conjunto difuso A em um universo X pode ser representado por:

$$ A = \bigcup_{\alpha \in \Lambda(A)} \alpha \cdot \left({}^{\alpha}A \right) $$

onde $\Lambda(A)$ é o conjunto dos níveis do conjunto difuso A, isto é, $\Lambda(A) := \{\alpha \mid A(x) = \alpha$ para um $x \in X\}$, $\bigcup$ é a união difusa e ${}^{\alpha}A$ representa a função característica do corte-α de A.

Exemplo 2.1: Seja $X = [0,100]$ e considerando um conjunto difuso A definido por:

$$ A(x) = \begin{cases} 0 & , se\ x < 60.5 \quad ou\ x > 80.5 \\ \frac{2}{15}x - \frac{121}{15} & , se\ 60.5 \leq x \leq 68 \\ 1 & , se\ 68 \leq x \leq 69.5 \\ -\frac{1}{11}x + \frac{80.5}{11} & , se\ 69.5 \leq x \leq 80.5 \end{cases} $$

então calcula-se como corte-α: ${}^{\alpha}A = [60.5 + 7.5\alpha;\ 80.5 - 11\alpha]$, para $\alpha > 0$, e ${}^{0}A = [0,100]$.

Exemplo 2.2: Seja C um conjunto difuso no universo finito $X = \{x_1,x_2,x_3,x_4,x_5\}$ definido por:

$$C = 0.2/x_1 + 0.4/x_2 + 0.8/x_3 + 0.4/x_4 + 1/x_5 ,$$

determina-se como conjunto de niveis $\Lambda(C) = \{0.2,0.4,0.8,1\}$ e como cortes-α:

$$^{0.2}C = \{x_1,x_2,x_3,x_4,x_5\};\ ^{0.4}C = \{x_2,x_3,x_4,x_5\};\ ^{0.8}C = \{x_3,x_5\};\ ^{1}C = \{x_5\}.$$

Pelo teorema de decomposição tem-se:

$$C = (0.2 \cdot {}^{0.2}C) \cup (0.4 \cdot {}^{0.4}C) \cup (0.8 \cdot {}^{0.8}C) \cup (1 \cdot {}^{1}C).$$

2.2.2. Relações difusas

Relações representam a presença ou ausência de associações, interações ou interconecções entre elementos de dois ou mais conjuntos [KLI 95]. As relações difusas generalizam este conceito, permitindo diversos graus de associação, interação ou interconecção entre os elementos. Relações clássicas podem ser identificadas com suas funções características, que atribuem o valor *1* aos elementos da relação e o valor *0* aos elementos do complemento. Semelhante à generalização de conjuntos, as relações difusas podem ser definidas através da generalização das funções características, permitindo um grau de pertinência na relação, isto é, sejam X_i, $1 \leq i \leq n$, conjuntos clássicos e seja $X_1 \times X_2 \times \ldots \times X_n$ o produto Cartesiano representando o universo; uma relação difusa R é definida como uma função

$$R : X_1 \times X_2 \times \ldots \times X_n \rightarrow [0,1].$$

Relações binárias têm uma importância especial entre as relações n-dimensionais, pois, de alguma forma, elas podem ser consideradas uma generalização de funções matemáticas. Ao contrário das funções, uma relação binária $R(X,Y)$ pode associar a cada elemento em X mais do que um elemento em Y. Considerando um universo finito, $X = \{x_1,x_2, \ldots ,x_n\}$; $Y = \{y_1,y_2, \ldots ,y_m\}$, pode-se identificar a relação binária $R(X,Y)$ com uma matriz R, onde os elementos $R_{i,j}$, $1 \leq i \leq n$; $1 \leq j \leq m$, indicam o grau de pertinência de (x_i,y_j) em $R(X,Y)$.

Como para funções, define-se para as relações binárias a inversa. A inversa de uma relação binária, difusa $R(X,Y)$ é uma relação difusa em $Y \times X$, $R^{-1}(Y,X)$, definido por

$$R^{-1}(y,x) := R(x,y), \text{ para todo } x \in X \text{ e todo } y \in Y.$$

As relações binárias difusas $R(X,X)$ definidas sobre um único conjunto X são também classificadas segundo diferentes propriedades características, como:

Reflexividade	$R(x,x) = 1$, para todos os $x \in X$
Simetria	$R(x,y) = R(y,x)$, para todos os $x, y \in X$
(max-min) Transitividade	$R(x,z) \geq \max_{y \in X} \min[R(x,y),R(y,z)]$, para todos os $(x,z) \in X^2$

Como no caso das operações básicas sobre conjuntos, existem várias possibilidades de generalizar a composição de funções. Uma dessas generalizações, baseada nas operações padrão, é a composição padrão, também chamada max-min composição. Sejam *P(X,Y)* e *Q(Y,Z)* duas relações binárias difusas; a composição padrão dessas duas relações, *P(X,Y)* ° *Q(Y,Z)*, é uma relação binária difusa, *R(X,Y)*, em *XxZ* definido por

$$R(x,z) = [P \circ Q](x,z) := \max_{y \in Y} \min[P(x,y), Q(y,z)] \text{ , para todo } (x,z) \in XxZ .$$

Exemplo 2.3: Sejam *P* e *Q* duas relações difusas

$$P = \begin{pmatrix} 0.3 & 0.5 & 0.8 \\ 0 & 0.7 & 1 \\ 0.4 & 0.6 & 0.5 \end{pmatrix} \qquad Q = \begin{pmatrix} 0.9 & 0.5 & 0.7 & 0.7 \\ 0.3 & 0.2 & 0 & 0.9 \\ 1 & 0 & 0.5 & 0.5 \end{pmatrix}$$

e $R = P \circ Q$ a max-min composição de *P* e *Q*. Os valores de *R* calculam-se como:

$$R_{1,1} = max[min(0.3,0.9), min(0.5,0.3), min(0.8,1)] = 0.8$$

$$R_{1,2} = max[min(0.3,0.5), min(0.5,0.2), min(0.8,0)] = 0.3, \text{ etc}$$

Considerando todas as combinações tem-se como resultado:

$$R = P \circ Q = \begin{pmatrix} 0.8 & 0.3 & 0.5 & 0.5 \\ 1 & 0.2 & 0.5 & 0.7 \\ 0.5 & 0.4 & 0.5 & 0.6 \end{pmatrix}$$

2.2.3. Lógica difusa

Lógica é o estudo de métodos e princípios de raciocínio em todas as formas possíveis. A lógica clássica trata de proposições que podem ter valores de verdade ou 'verdadeiro' ou 'falso'. Cada proposição tem seu oposto, a negação, e é exigido que eles assumam valores de verdade opostos. Essa suposição básica da lógica clássica foi questionada desde Aristóteles.

O primeiro a considerar lógicas de *n* valores, $n \geq 2$, foi Lukasiewicz no início dos anos 30 deste século [KLI 95]. O caso extremo representa uma lógica cujos valores de verdade podem assumir todos os números reais no intervalo *[0,1]*. Essa lógica com infinitos valores é geralmente chamada lógica padrão de Lukasiewicz L_1.

2.2.3.1. Primitivas padrão da lógica difusa

Lógica proposicional e a teoria de conjuntos finitos representam álgebras Booleanas, portanto são isomorfos identificando as operações: união e disjunção;

intersecção e conjunção; complemento e negação. Da mesma forma define-se as primitivas padrão da lógica difusa de tal maneira que a lógica L_1 é isomorfa à teoria de conjuntos finitos difusos baseada nas operações padrão difusas. Portanto define-se:

negação	$V(\neg A) = 1 - V(A)$
conjunção	$V(A \wedge B) = min(V(A), V(B))$
disjunção	$V(A \vee B) = max(V(A), V(B))$

onde A e B representam proposições; $\neg$, $\wedge$, $\vee$ representam as operações de negação, conjunção e disjunção, respectivamente, e V é uma função de verdade.

2.3. Considerações finais do capítulo

O conceito de variáveis lingüísticas introduzido por Zadeh [ZAD 75], traz uma ligação entre linguagem natural e teoria difusa. Variáveis lingüísticas diferem de variáveis numéricas pelos valores, os quais não são valores numéricos, mas palavras ou frases. Uma variável lingüística é caracterizada pelo seu conjunto de termos, i.e. o conjunto de possíveis valores lingüísticos que a variável pode obter. Termos ou valores lingüísticos são tipicamente adjetivos ou combinações de adjetivos com modificadores ("muito", "pouco","mais ou menos", etc) e conectores ("e", "ou", "ou ... ou", etc) que tem associado um conjunto difuso descrevendo o sentido do termo. Por exemplo, a variável lingüística 'idade' poderia ter como possíveis valores 'jovem', 'muito jovem', 'não muito jovem', 'velho', etc. Zadeh [ZAD 75] interpreta os conjuntos difusos associados aos valores lingüísticos, por exemplo 'jovem', como restrições difusas nos valores de uma variável base, neste exemplo 'idade', que descrevem a compatibilidade de possíveis valores da variável base, por exemplo 22, 33, com os valores lingüísticos.

O modelo de recuperação de documentos apresentado no capítulo 4 utiliza tanto processamento de linguagem natural quanto a teoria difusa. Mas a ligação entre linguagem natural e teoria difusa, considerada no capitulo 4, é diferente do conceito das variáveis lingüísticas.

3. RECUPERAÇÃO DE DOCUMENTOS

Segundo Zemankova & Kandel [ZEM 85], a pesquisa em recuperação de informações imprecisas é dividida em duas áreas: recuperação de documentos e recuperação de fatos. A recuperação de documentos refere-se ao objetivo de encontrar referências bibliográficas de documentos. A idéia básica atrás da recuperação de documentos é encontrar documentos que contenham informações tanto conveniente quanto possíveis para a consulta de um usuário. A recuperação de fatos é a recuperação de dados dependendo de uma concordância entre valores de atributos da consulta e valores de itens de dados de uma base de dados.

3.1. Um modelo genérico de recuperação de documentos

Na literatura pode-se encontrar vários modelos genéricos de recuperação de documentos como, por exemplo, o modelo de Radecki [RAD 81], o modelo de Zenner, Caluwe e Kerre [ZEN 85] e o modelo de Murai, Miyakoshi e Shimbo [MUR 89]. Todos estes modelos consideram um conjunto de documentos, um conjunto de descritores, um conjunto de consultas e uma função de recuperação. Como exemplo, apresenta-se aqui o modelo genérico de Murai, Miyakoshi e Shimbo [MUR 89]. Assim, um modelo de um sistema de recuperação de documentos é uma 9-tupla[1]

$$(D,D',X,C,C',\psi,\chi,\rho,S)$$

onde

- D é um conjunto de documentos;
- X é um conjunto de descritores; descritores podem ser palavras-chave ou outros símbolos que servem para caracterizar o conteúdo de documentos;
- D' é um conjunto de representações dos documentos; isto pode ser para cada documento um subconjunto de X indicando os descritores do documento, ou para cada documento um conjunto difuso em X indicando a relevância de cada descritor para o documento
- C é um conjunto de consultas;
- C' é um conjunto de representações das consultas; estas podem ser, por exemplo, representações em linguagem formal de consultas em linguagem natural;
- S é o conjunto de valores de status da recuperação, por exemplo $\{0,1\}$ no caso booleano e $[0,1]$ no caso difuso,
- ψ: $D \rightarrow D'$ é uma função de indexação, que atribui aos documentos suas representações;

[1] Em [MUR 89], o modelo genérico é definido como uma 7-tupla, considerando D' e C' somente na explicação da função de indexação e da função de consultas.

- χ: $C \rightarrow C'$ é uma função de consultas, que transforma consultas de usuários em representações que o sistema pode interpretar; isto pode ser, por exemplo, a tradução de consultas em linguagem natural para uma linguagem formal;
- ρ: $C' \times D' \rightarrow S$ é a função de recuperação, que associa a cada representação de uma consulta e a cada representação de um documento um valor de status de recuperação.

Neste modelo, a resposta, R, de uma consulta c é um conjunto de duplas (d,s) em $D \times S$

$$R = \{(d,s) \in D \times S \mid \rho(\chi(c), \psi(d)) = s\}$$

Muitos autores, por exemplo [MIY 86; 89; 90], [ZEN 85], [MUR 88], etc, identificam os conjuntos de documentos D e o conjunto das consultas C com os conjuntos das representações D' e C', e consideram somente um modelo (D,X,C,ρ,S). O valor de status de recuperação pode servir para decidir quais são os documentos retornados para o usuário e definir uma ordem de apresentação dos documentos. Muitos sistemas consideram $S = \{0,1\}$, diferenciando somente entre os fatos de o documento ser retornado ou não. Radecki [RAD 81] considera, neste caso, um modelo (D,X,C,ρ') com uma função de recuperação $\rho': C \rightarrow 2^D$. Assim, uma resposta, R, de uma consulta $c \in C$ pode ser considerada como conjunto

$$R = \rho'(c) = \{d \in D \mid \rho(c,d) \geq \alpha\}$$

onde α é um threshold que determina se um documento é retornado ou não.

3.2. Thesaurus

O primeiro thesaurus para a recuperação de documentos foi publicado em 1959 nos Estados Unidos. Em 1970 a UNESCO publicou as primeiras diretrizes para a construção de thesauri monolíngües e em 1976 apareceram as primeiras diretrizes da UNESCO para a construção de thesauri multilíngües [LAN 87].

Segundo Lancaster [LAN 87] a construção de um thesaurus tradicional se divide em quatro etapas fundamentais:

- a coleta dos termos indexados de forma controlada
- ordenação dos termos
- organização dos termos estabelecendo relações entre os mesmos
    - hierárquicas
        - ⇒ todo/ parte
        - ⇒ gênero/espécie
    - associativas
- apresentação e impressão

A coleta dos termos determina sobre o que tratam os documentos e faz uma seleção de termos de indexação para representar o conteúdo dos documentos. O conjunto de todos os termos de indexação utilizados é denominado vocabulário ou

linguagem de indexação. Segundo Lancester é preferível que este vocabulário seja controlado para ter melhores resultados na busca de algum tópico (abordagem antiga). Mas, isto requer que indexadores e usuários utilizem a mesma linguagem.

Para a coleta dos termos indexados foram desenvolvidos métodos automáticos: por exemplo Robredo [ROB 82], Salton e Smith [SAL 89], Grefenstette [GRE 92]. Também para a construção das relações entre os termos foram desenvolvidas diversas técnicas automáticas, tanto não difusas: por exemplo Lukashevich [LUK 95] quanto difusas, por exemplo Miyamoto, Miyake e Nakayama [MIY 83] ou Larsen e Yager [LAR 93].

3.2.1 Construção de um thesaurus difuso

Como thesaurus[2] difuso entendemos uma relação reflexiva, difusa entre os termos indexados. Em [MIY 83;86;90] encontram-se três fórmulas para a construção destas relações baseadas nas co-ocorrências dos termos indexados. Duas destas são relações difusas de proximidade representando relações associativas entre termos indexados e uma é relação difusa de inclusão, representando relações hierárquicas entre termos indexados.

Seja $D = \{d_1, d_2, \ldots, d_m\}$ um conjunto de documentos, seja $X = \{x_1, x_2, \ldots, x_n\}$ um conjunto de termos indexados e seja $I: X \times D \rightarrow \{0,1\}$ o índice que associa documentos aos termos indexados. $I(x_i,d_j) = 1$ se x_i é termo indexado do documento d_j, e $I(x_i,d_j) = 0$ se x_i não faz parte dos termos indexados associado ao documento d_j. As três relações difusas s_1, s_2, t definidas abaixo servem para a construção de um thesaurus difuso [MIY 83;86;90].

- relação de inclusão:

$$t(x_i, x_j) = \frac{\sum_{k=1}^{m} min\left(I(x_i, d_k), I(x_j, d_k)\right)}{\sum_{k=1}^{m} I(x_i, d_k)}$$

uma relação reflexiva, mas não necessariamente simétrica, que tem o valor 1 se x_i é somente termo indexado de um documento, quando x_j também é termo indexado deste documento;

[2] Miyamoto e Nakayama [MIY 86] diferenciam entre thesaurus e pseudo-thesaurus. Os autores consideram no mesmo artigo relações de proximidade, um thesaurus difuso como outros autores também. Miyamoto [MIY 89;90] considera tanto relações de inclusão quanto relações de proximidade thesauri difusos.

- relações de proximidade

$$s_1(x_i, x_j) = \frac{\sum_{k=1}^{m} min\left(I(x_i, d_k), I(x_j, d_k)\right)}{\sum_{k=1}^{m} max\left(I(x_i, d_k), I(x_j, d_k)\right)}$$

$$s_2(x_i, x_j) = min\left(t(x_i, x_j), t(x_j, x_i)\right)$$

que são reflexivas e simétricas.

3.3. Exemplos de modelos

Segundo Zemankova & Kandel [ZEM 85] os métodos de busca mais utilizados em recuperação de informações são do tipo identidade exata ou recuperação aproximada. No caso de recuperação aproximada, os candidatos mais próximos são recuperados, utilizando um critério de proximidade apropriado. Medidas de proximidade utilizadas são, por exemplo, funções de distância como a distância Euclideana, ou relações de similaridade. Segundo Zemankova e Kandel [ZEM 85], muitos destes métodos de recuperação aproximada são somente de natureza teórica. Alguns foram implementados em sistemas experimentais, mas só poucos são utilizados em sistemas comerciais de recuperação de informações. Segundo Russel e Norvig [RUS 95] considerou-se antigamente modelos booleanos e modelos baseados em processamento de linguagemnatural, independemente. Os modelos booleanos desenvolveram-se para modelos em espaços vetoriais. Hoje em dia, considera-se modelos híbridos que combinam as duas abordagens.

Existem vários tipos de modelos como por exemplo sistemas especialistas, (CanSearch) Pollitt [POL 87]); redes neuronais: Belew [BEL 89], modelos lógicos, difusos ou não-difusos: Murai, Miyakoshi & Shimbo [MUR 88], Radecki [RAD 81], sistemas que utilizam grafos difusos ou não-difusos: Eastman & Weiss [EAS 78]; Nomoto, Wakayama, Kirimoto, Ohashi & Kondo [NOM 90], sistemas que utilizam processamento de linguagem natural, Hess [HES 92], etc...

3.3.1. O modelo de Miyamoto

O modelo de Miyamoto [MIY 89; 90] supõe que cada documento é caracterizado por um conjunto de palavras-chave ou outros símbolos. O modelo de Miyamoto envolve dois conjuntos clássicos.

- um conjunto de termos indexados (palavras chaves): $X = \{x_1, x_2, \ldots, x_n\}$
- um conjunto de documentos : $D = \{d_1, d_2, \ldots, d_m\}$

Esses conjuntos podem ser modificados por inclusão ou exclusão de documentos, mas para cada consulta estes dois conjuntos são fixos. Além disso, considera-se duas relações difusas:

- a relação de relevância dos termos indexados para os documentos,

$$V : D \times X \rightarrow [0,1],$$

especificando para cada termo indexado e para cada documento o grau de relevância do termo indexado para o documento;

- uma relação reflexiva, um thesaurus difuso:

$$T : X \times X \rightarrow [0,1],$$

que expressa o grau de compatibilidade do significado entre dois termos indexados. Esta relação deve ser reflexiva, pois cada termo indexado é compatível consigo.

Uma consulta neste sistema é um conjunto difuso C em X. Através do thesaurus difuso esta consulta é estendida, determinando os termos indexados com significado compatível com os termos da consulta. Segundo Klir & Yuan [KLI 95], utiliza-se geralmente para esta consulta aumentada, $C_s = T \circ C$, a max-min composição, i.e.

$$C_s(x_i) = \max_{x \in X} \min\left(T(x_i, x), C(x)\right), \text{para todo } x_i \in X.$$

A relevância dos documentos para os termos da consulta obtém-se através da composição da consulta aumentada com a relação de relevância dos termos indexados para os documentos relevantes:

$$R(d_j) = (V \circ C_s)(d_j) = \max_{x \in X} \min\left(V(d_j, x), C_s(x)\right), \text{para todo } d_j \in D,$$

Para limitar os documentos que são apresentados ao usuário este sistema considera a existência de um filtro. Um filtro poderia ser um corte-α de R, onde α, $0 \leq \alpha \leq 1$, é um parâmetro especificado pelo usuário,

$$^{\alpha}R = \{d \in D \mid R(d) \geq \alpha\}.$$

A resposta poderia ser uma listagem dos documentos em $^{\alpha}R$ em ordem decrescente dos valores de pertinência em R.

Comparando este modelo de Miyamoto com o modelo genérico de Murai, Miyakoshi e Shimbo [MUR 89] pode-se observar que neste modelo a função de recuperação pode ser considerada uma composição do thesaurus difuso com a relação de relevância dos termos indexados para os documentos, i.e.

$$\rho(C, \bullet) = V \circ T \circ C, \text{ para uma consulta } C.$$

O modelo considera ainda a aplicação de um filtro *F*, que determina quais são os documentos que compõem a resposta e em que forma esta resposta é apresentada ao usuário. Miyamoto [MIY 90] representa este modelo como na figura 3.1, em que uma consulta passa por três módulos antes de retornar a resposta.

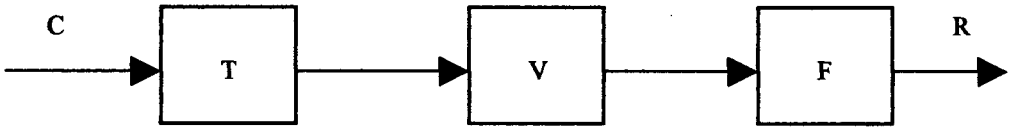


Figura 3.1: Representação esquemática do modelo de Miyamoto.

Segundo Miyamoto [MIY 90] outros tipos de filtros podem servir, por exemplo, para que o sistema retorne somente documentos de uma determinada área de interesse ou somente documentos mais recentes.

Em seguida é apresentado um exemplo de Klir & Yuan [KLI 95] para o sistema de Miyamoto (exemplo 3.1). Logo depois apresenta-se um método para a construção da relação de relevância dos termos indexados para os documentos para este modelo através de um thesaurus difuso.

Exemplo 3.1:

O exemplo considera somente seis termos indexados relevantes para uma consulta que envolve três destes termos. Sejam

x_1 = fuzzy logic, x_2 = fuzzy relation equations

x_3 = fuzzy modus ponens, x_4 = approximate reasoning

x_5 = max-min composition, x_6 = fuzzy implication

e seja a representação vetorial da consulta *C* dada por:

$$C = \begin{array}{c} \underline{\quad x_1 \quad x_2 \quad x_3 \quad} \\ (1 \quad 0.4 \quad 0.1) \end{array}$$

Ainda supõe-se que a parte relevante do thesaurus difuso (as linhas restritas ao suporte de *C* e as colunas resultantes de *T*, restritas às colunas não zeros) é dada por:

$$T^{-1} = \begin{array}{c} \begin{array}{cccccc} \underline{x_1} & \underline{x_2} & \underline{x_3} & \underline{x_4} & \underline{x_5} & \underline{x_6} \end{array} \\ \begin{pmatrix} 1 & 0.2 & 1 & 1 & 0.5 & 1 \\ 0.2 & 1 & 0.1 & 0.7 & 0.9 & 0 \\ 1 & 0.4 & 1 & 0.9 & 0.3 & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ x_1 \\ x_2 \\ x_3 \end{array}$$

Considerando este thesaurus, pode-se determinar a consulta aumentada C_S como:

$$\begin{matrix} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix}$$

$$C_S = T \circ C = \begin{pmatrix} 1 & 0.4 & 1 & 1 & 0.5 & 1 \end{pmatrix}$$

Seja a parte relevante (as linhas restritas ao suporte de C_S e as colunas resultantes restritas às colunas não-zeros) da relação de relevância dos termos indexados para os documentos dada pela matriz:

$$\begin{matrix} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \end{matrix}$$

$$V^{-1} = \begin{pmatrix} 0.2 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0.3 & 0 & 0.4 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.8 & 0 & 0.4 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.7 & 0.5 \\ 1 & 0 & 0.5 & 0 & 0 & 0.6 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.2 & 0 & 1 & 0 & 0 & 0.5 \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{matrix},$$

onde d_1, d_2, ..., d_{10} são os únicos documentos que se relacionam com os termos indexados considerados. Portanto tem-se como documentos recuperados:

$$\begin{matrix} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \end{matrix}$$

$$R = V \circ C_S = \begin{pmatrix} 0.5 & 1 & 1 & 0.3 & 0.4 & 0.5 & 1 & 0.9 & 0.7 & 0.5 \end{pmatrix}$$

Considerando ainda um corte-α de 0.7 determina-se

$${}^{0.7}R = \{d_2, d_3, d_7, d_8, d_9\}.$$

Colocando estes cinco documentos em ordem decrescente dos valores de pertinência em R tem-se como resposta:

$$d_2, d_3, d_7, d_8, d_9.$$

3.3.1.1 Construção da relação de relevância dos termos indexados para os documentos

Para o modelo de recuperação de documentos apresentado anteriormente, Miyamoto [MIY 89; 90] propôs um método de construção da relação de relevância dos termos indexados para os documentos, que já havia sido utilizado por Miyamoto e Nakayama [MIY 86] em um outro modelo.

Seja I um índice, isto é, uma relação binária não-difusa $I: X \times D \rightarrow \{0,1\}$ que associa documentos aos termos indexados. O método proposto por Miyamoto determina a relação de relevância dos termos indexados para os documentos $V: X \times D \rightarrow [0,1]$ como uma extensão difusa do índice não-difuso $I: X \times D \rightarrow \{0,1\}$ através de um thesaurus difuso preestabelecido $T: X \times X \rightarrow [0,1]$.

O grau de relevância de um termo indexado para um documento é considerado o máximo dos valores de pertinência no thesaurus difuso entre este termo indexado e as palavras-chave associadas ao documento, i.e. $V: X \times D \rightarrow [0,1]$ é definido como:

$$V(x_i, d_j) := \max_{x \in (I(\bullet, d_j))^{-1}(\{1\})} T(x_i, x)$$
$$= \max_{x \in X} \min\left(T(x_i, x), I(x, d_j)\right)$$

onde $(I(\cdot, d_j))^{-1}(\{1\}) = \{x \in X \mid I(x, d_j) = 1\}$ é o conjunto dos descritores do documento d_j.

3.3.2. O modelo de Bruza e van der Weide

O modelo de Bruza e van der Weide [BRU 91] é um modelo conceitual baseado em lógica que utiliza expressões indexadas para representar o conteúdo de documentos. Expressões indexadas são consideradas uma seqüência de termos e conectores. Os conectores são considerados como palavras que descrevem as relações entre os termos. A sintaxe das expressões indexadas é especificada da seguinte forma (BNF estendida):

Expressão → *Termo {Conector Expressão}**
Termo → *string*
Conector → *string*

Os termos correspondem a substantivos, adjetivos ou sintagmas nominais. Os conectores são basicamente restritos às preposições e um chamado conector nulo, que é representado pelo símbolo '•'. A figura 3.2 mostra alguns dos conectores permitidos e o tipo de relação que eles representam.

Conector	Tipo de relação	Exemplo
of	posse	castle of queen
	ação-objeto	pollination of crops
by	ação-agente	voting by students
in, on	posição	trees in garden
to, on,for, in	associação direta	attitudes to courses research on voting
with, •, and	associação	assistance with problems fruit • trees
as	equivalência	

Figura 3.2 Conectores e os tipos de relação associados (figura 1 em [BRU 91])

As possíveis ambigüidades de expressões indexadas são tratadas pela representação em árvores. Para mostrar como isto é feito Bruza e van der Weide consideram o exemplo 'attitudes to courses of students in universities'.

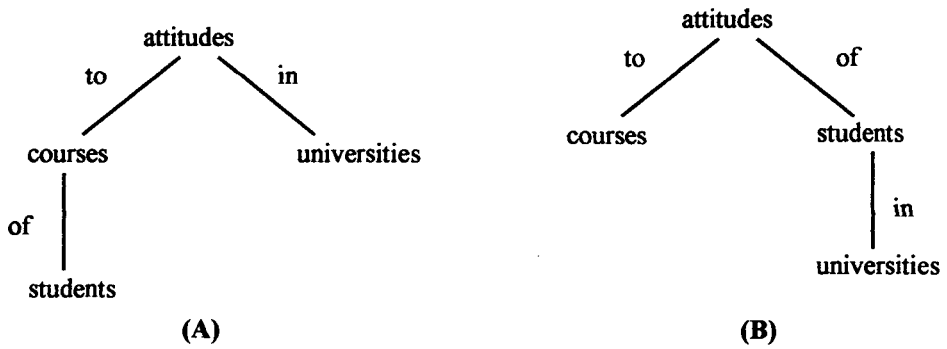


Figura 3.3: Representações de diferentes interpretações da expressão indexada 'attitudes to courses of students in universities' [BRU 91].

A questão é se as atitudes são atitudes de estudantes universitários em relação aos cursos (figura 3.3 **B**) ou atitudes em universidades em relação a cursos estudantis (figura 3.3 **A**).

Para a modelagem da recuperação, os autores introduzem o conceito da potência de interpretações de expressões indexadas, i.e. o conjunto de todas as sub-expressões que se consegue removendo somente folhas, ou seja todas as sub-árvores. A figura 3.4 mostra a potência da interpretação mostrada na figura 3.3 **A**.

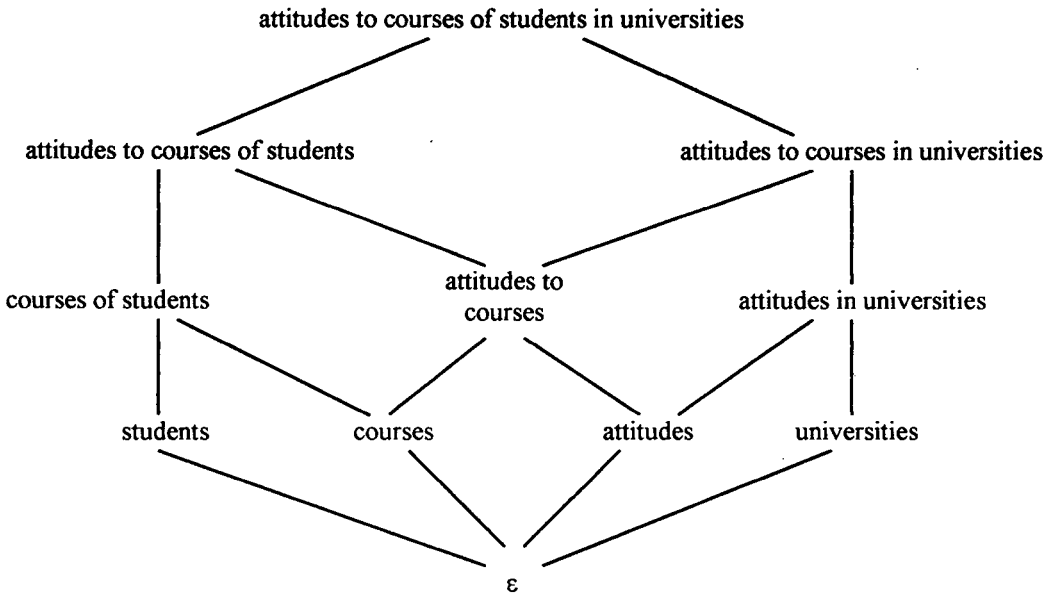


Figura 3.4: Potência da interpretação de 'attitudes to courses of students in universities' representada na figura 3.3 (A) (ε = string vazio) [BRU 91]..

Observa-se, por exemplo, que 'attitudes of students' não é sub-expressão. A recuperação é feita por comparação de todas as interpretações da expressão de consulta com as interpretações das expressões indexadas associadas aos documentos. Se, por exemplo, a expressão de consulta é uma sub-expressão de uma expressão indexada associada ao documento *d*, então *d* é retomado.

3.3.3. 'Stemming'

Sistemas de recuperação de documentos que utilizam 'Stemming' consideram termos com radicais em comum como relacionados. O 'Stemming' é utilizado tanto para a construção de thesauri quanto para a determinação da relação entre termos da consulta e termos indexados, bem como para a compressão do índice. Segundo Lopes [LOP 96] existem 4 tipos de 'Stemming':

- 'Table Lookup Stemming': Prevê a existência de uma tabela de radicais e compara termos indexados com os radicais da tabela. Um dos problemas deste método é a não existência de uma tabela de radicais aproximadamente completa.
- 'Stemming' por variedade de sucessores: Compara os termos com palavras de um corpus. Determina para cada possível radical de um termo a variedade de sucessores, isto é, o número de letras que seguem em uma palavra do corpus, este possível radical do termo. Determina através da variedade de sucessores um ponto de corte. Isto pode ser, por exemplo, o possível radical cuja variedade de sucessores representa um pico.

Exemplo 3.2:

Corpus: able, ape, fixable, read, reading, reads, readability, red, rope, ripe.
Termo: readable

Possíveis radicais	Variedade de sucessores (vs)	Letras que seguem
r	3	e,i,o
re	2	a,d
rea	1	d
read	3	i,s,' '
reada	1	b
readab	1	i
readabl	0	
readable	0	

Observa-se um pico *(*vs('rea') = *1* < *3* = vs('read') e vs('read') > vs('reada') =*1)* para o possível radical 'read'.

- 'Stemming' baseada em n-gramas: Estabelece uma medida de similaridade entre duas palavras comparando a soma dos n-gramas (sequência de n-letras consecutivas) das duas palavras com os n-gramas compartilhados.

Exemplo 3.3

termo1: statistics → st-ta-at-ti-is-st-ti-ic-cs;
digramas: at, cs, ic, is, st, ta, ti
termo2: statistical → st-ta-at-ti-is-st-ti-ic-ca-al;
digramas: al, at, ca, cs, ic, st, ta, ti

digramas compartilhados: at, cs, ic, st, ta, ti

$$similaridade = \frac{2 * |digramas\ compartilhados|}{|digramas_{termo1}| + |digramas_{termo2}|} = \frac{2 * 6}{7 + 8} = 0.8 ,$$

onde |X| denota o número de elementos do conjunto X.

- 'Stemming' por remoção de afixos: Prevê a existência de uma tabela de afixos. Descobre possíveis radicais de termos por comparação do início e/ou final dos termos com os afixos. Remove afixos segundo regras preestabelecidas. Por exemplo, remove o maior sufixo que coincide com o final do termo.

3.4. Considerações finais do capítulo

Segundo Klir & Yuan [KLI 95], a recuperação de documentos baseada na teoria de conjuntos difusos vem sendo reconhecida como mais realística do que os métodos clássicos de recuperação de documentos. Existem diversos modelos difusos de recuperação de documentos apresentados na literatura mas, segundo Klir & Yuan [KLI 95], o modelo de Miyamoto representa o único tratamento compreensível deste assunto. Por outro lado o modelo de Miyamoto supõe a existência de um vocabulário de consulta controlado, preestabelecido, sendo que isto representa uma restrição considerável no caso de usuários não treinados.

O modelo de Bruza & van der Weide [BRU 91] é mais flexível na linguagem de consulta. Os testes mostraram resultados semelhantes a outros sistemas no que diz respeito à eficiência. No entanto, este apresenta um algoritmo de recuperação complexo. Além disso, segundo os autores, ainda deveria ser investigado se o modelo é computacionalmente tratável.

Quanto ao emprego de 'Stemming', este apresenta a vantagem de conferir maior flexibilidade na formulação das consultas. No entanto, pode levar a erros na determinação dos radicais devido à falta de uma lista relativamente completa de radicais ('Table Lookup Stemming'), à qualidade do corpus ('Stemming' por variedade de sucessores) ou a regras não apropriadas de remoção de afixos.

4. O MODELO PROPOSTO

Este trabalho propõe um modelo de recuperação de documentos, baseado em um algoritmo de reconhecimento de radicais e sufixos. Este modelo é uma extensão do modelo de recuperação de informações de Miyamoto [MIY 89; 90]. O conceito de similaridade entre palavras baseia-se, por um lado, em um thesaurus difuso como no sistema de Miyamoto e, por outro lado, em comparações entre radicais determinados pelo sistema de reconhecimento de radicais e sufixos. O modelo proposto baseia-se na comparação de radicais detectando possíveis similaridades lexicais entre palavras com significado idêntico ou semelhante, enquanto o thesaurus difuso serve para detectar significados semelhantes (similaridades semânticas).

Comparado com o modelo de Miyamoto, este modelo é composto de quatro módulos, um a mais do que no modelo de Miyamoto (figura 4.1).

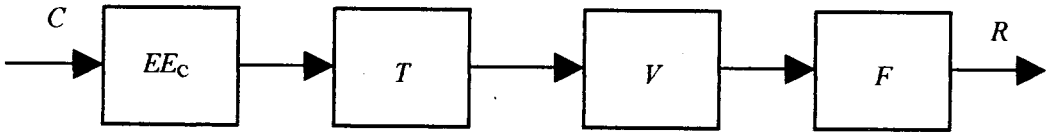


Figura 4.1: Representação esquemática do modelo proposto.

O módulo EE_C representa um ajuste lexical da consulta. Os outros módulos representam, como no modelo de Miyamoto, um thesaurus, *T*, uma relação entre descritores e documentos, *V*, e um filtro, *F*. Para a construção do ajuste lexical, EE_C, considera-se primeiro uma relação difusa entre palavras, que é estabelecida através de um sistema de reconhecimento de radicais e sufixos.

4.1. A Relação palavra/palavra

O sistema de reconhecimento de radicais e sufixos é basicamente o sistema apresentado em [STO 96] (anexo 1). Como já indicado no capítulo 2.1.4, os termos sufixo e radical devem ser interpretados neste contexto como conceitos computacionais. Sufixo poderia ser tanto uma terminação de flexão, quanto terminações nominais ou verbais ou combinações destas terminações. O termo radical deve ser entendido como o restante da palavra.

O sistema de reconhecimento de radicais e sufixos determina para cada palavra uma tabela contendo possíveis pares radical/sufixo. Para cada par radical/sufixo é calculado um grau de certeza, um valor entre *0* e *1*, baseado na combinação das certezas obtidas para o radical e o sufixo. A figura 4.2 mostra uma tabela associada à palavra "difuso" com possíveis pares radical/sufixo e certezas associadas.

Posição do corte	Radical	Sufixo	$certeza_{par}$
1	d	ifuso	0.0
2	di	fuso	0.1
3	dif	uso	0.3
4	difu	so	0.1
5	difus	so	0.8
6	difuso		0.3

Figura 4.2 Todas as possibilidades de pares radical/sufixo da palavra "difuso" com valores de certeza dos pares radical/sufixo ($certeza_{par.}$).

Formalmente as tabelas associadas às palavras podem ser interpretadas como uma relação difusa, em que a cada palavra e a cada possível radical é atribuído um valor de pertinência, a certeza do par que forma com este radical a palavra. Seja esta relação difusa entre palavras e radicais definida por

$$PR : \Sigma^* \times \Sigma^* \rightarrow [0,1]$$

$$PR(p,r) = \begin{cases} certeza_{par(r,s)} & se\ o\ par(r,s)\ aparece\ na\ tabela\ da\ palavra\ p \\ 0 & nos\ outros\ casos \end{cases}$$

onde $\Sigma = \{a,b, \ldots ,z\}$ é o alfabeto da língua portuguesa, e Σ^* é o fechamento de Σ pela concatenação, i.e. o conjunto de strings finitos. Para cada palavra esta relação descreve a certeza do sistema sobre os possíveis radicais de cada palavra. Da mesma forma a inversa desta relação, PR^{-1}, descreve a certeza do sistema sobre a relação radical/palavra, dado um determinado radical. Combinando estas duas relações, define-se uma relação entre palavras, *PP*, que descreve a certeza do sistema de que duas palavras têm o mesmo radical.

$$PP : \Sigma^* \times \Sigma^* \rightarrow [0,1]$$

$$PP(p,p) = 1 \quad e$$

$$\begin{aligned} PP(p_1,p_2) &= (PR \circ PR^{-1})(p_1,p_2) \\ &= \max_{r \in \Sigma^+} \min\left(PR(p_1,r), PR^{-1}(r,p_2)\right) \\ &= \max_{r \in \Sigma^+} \min\left(PR(p_1,r), PR(p_2,r)\right) \\ &= \max_{r \in (rad(p_1) \cap rad(p_2))} \min\left(PR(p_1,r), PR(p_2,r)\right) \end{aligned}$$

onde p, p_1 e p_2 representam palavras, $rad(p)$ representa o conjunto de radicais da palavra p na tabela associada, e considerando, por convenção, que o máximo sobre um conjunto vazio é *0*.

4.2. Análise de expressões

Como no modelo de Bruza e van der Weide [BRU 91], as expressões são interpretadas como seqüências de termos e conectores. Supõe-se que os termos são palavras da classe aberta e os conectores são palavras da classe fechada. A semelhança entre expressões é interpretada como semelhança entre os termos das expressões e a semelhança das relações entre os termos é estabelecida por preposições e contrações de artigo e preposição. A figura 4.3 mostra os conectores e as relações considerados.

Relações	Conectores	Exemplos
posse, material, conteúdo	de, do, da, dos, das com	cama de ferro, jarra de água, nariz do Pedro, casa do João, menina com cabelo comprido.
ação objeto	de, do, da, dos, das	invasão da cidade, eleição do presidente.
ação agente	de, do, das, dos, a, à, ao, às, aos por, pelo, pela, pelos pelas, via	invasão dos bárbaros, eleição do presidente por deputados, eleição do presidente pelos deputados, eleição do presidente via deputados.
posição, lugar, direção	de, do, da, das, dos, a, à, ao, às, aos, em, no, na, nas, nos, dentro, por, para, sobre.	as chegadas por diversos caminhos, a faca na mesa, o caminho para a cidade o rei sobre nós
associação	de, do, da, das, dos, a, à, ao, às, aos, por, pelo, pela, pelas, pelos, •, sobre, em, com, conforme, segundo, contra.	o trabalho sobre matemática, o trabalho em inglês, o trabalho com alguma pessoa, o trabalho para o professor, o voto contra a reeleição, o voto pela reeleição.
equivalência	como, de,	cabelo de milho

Figura 4.3: Conectores e as relações que eles podem representar.

Para determinar a semelhança de duas expressões, e_1 e e_2, cada expressão é transformada em uma seqüência de palavras. Desta seqüência são removidas as palavras da classe fechada que não são conectores. A palavra 'a' que pode ser tanto conector (preposição) quanto não conector (artigo) é somente removida se a palavra vizinha da esquerda é um conector. Entre duas palavras abertas sem conector no meio inclui-se o conector nulo, '•'. Semelhante ao modelo de Bruza e van der Weide [BRU 91], considera-se para cada expressão um grafo direcionado rotulado, que tem as palavras abertas como nodos e os conectores como rótulos dos arcos. A figura 4.4 mostra os grafos para duas expressões.

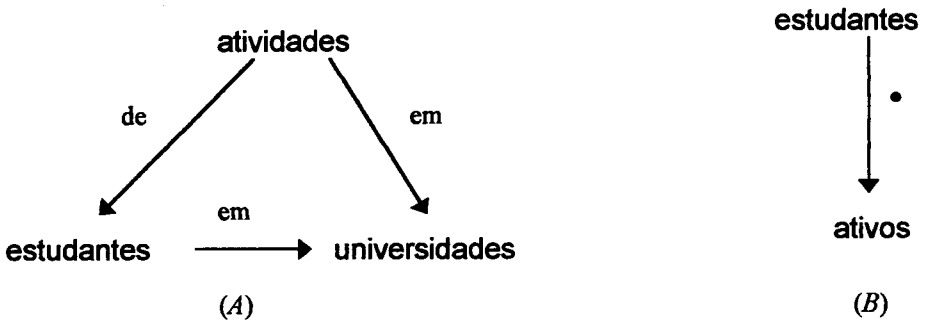


Figura 4.4: Dois exemplos de grafos associados a expressões: (*A*) grafo g_1 para a expressão e_1 = 'atividades de estudantes em universidades'; (*B*) grafo g_2 para a expressão e_2 = 'estudantes ativos'.

Para a determinação da semelhança entre as expressões, compara-se cada expressão com os caminhos (não considerando a direção) no grafo estabelecido para a outra expressão.

A seguir considera-se somente a expressão e_1 e o grafo g_2 determinado para a expressão e_2. Na comparação entre e_1 e os caminhos de g_2 considera-se somente os caminhos que incluem um número de palavras abertas igual ou menor do que em e_1. À cada caminho em g_2 atribui-se um valor de semelhança determinado como mínimo das comparações das palavras abertas do caminho e de e_1 na mesma posição através da relação entre palavras PP. Deste valor das comparações subtrai-se algumas penalidades. Atribui-se penalidades nos seguintes casos:

- o caminho percorre um arco em direção errada;
- o caminho percorre um arco em direção corrreta, mas o conector na mesma posição em e_1 não pode representar a mesma relação;
- o caminho contém menos palavras abertas do que a expressão e_1; atribui-se uma penalidade para cada palavra aberta a mais na expressão e_1 do que no caminho considerado em g_2.

Determina-se o caminho em g_2 com o maior valor de semelhança com a expressão e_1 e , da mesma forma, o caminho em g_1 com o maior valor de semelhança com a expressão e_2. O mínimo desses dois valores é estabelecido como o valor de semelhança entre as duas expressões, $EE(e_1,e_2)$.

Exemplo 4.1:

Considerando-se as expressões e_1 e e_2 da figura 4.4. Supõe-se

- que PP('atividades','ativos') = 0.8 e nos outros casos $PP(x,y) = 0$, se $x \neq y$;
- penalidade de 0.1 para o percurso de um arco em direção errada e penalidade de 0.2 para cada palavra da classe aberta ainda não considerada.

Determina-se a semelhança dos caminhos em g_1 com a expressão e_2 como

- estudantes ⇒ 1 - 0.2 = 0.8;
- estudantes → atividades ⇒ min(1;0.8) - 0.1 = 0.7;
- igual a 0 para os outros caminhos;

Determina-se a semelhança dos caminhos em g_2 com a expressão e_1 como

- ativos ⇒ 0.8 - 0.2 - 0.2 = 0.4;
- ativos → estudantes ⇒ min(0.8;1) - 0.1 - 0.2= 0.5;
- igual a 0 para os outros caminhos;

Portanto tem-se $EE(e_1,e_2)$ = min(0.8;0.5) = 0.5.

4.3. Recuperação difusa

O sistema difuso de recuperação de informações considerado aqui é uma extensão do modelo de Miyamoto [MIY 89; 90]. Como no modelo de Miyamoto, cada documento é caracterizado por um conjunto de descritores, aqui considerados expressões.

Este modelo de recuperação de documentos envolve dois conjuntos clássicos:

- um conjunto de documentos $D = \{d_1, d_2, \ldots, d_m\}$;
- um conjunto de descritores (expressões) $X = \{x_1, x_2, \ldots, x_n\}$

Esses conjuntos podem ser modificados por inclusão ou exclusão de documentos, mas para cada consulta estes dois conjuntos são considerados fixos. Além disso, considera-se duas relações difusas:

- a relação de relevância dos descritores para os documentos:

$$V : D \times X \rightarrow [0,1],$$

especificando para cada descritor e para cada documento o grau de relevância do descritor para o documento;

- uma relação reflexiva, o thesaurus difuso entre os descritores:

$$T : X \times X \rightarrow [0,1],$$

que expressa o grau de compatibilidade do significado entre dois descritores.

Uma consulta neste sistema é um conjunto difuso C de expressões, i.e.

$$C = c_1/t_1 + c_2/t_2 + \ldots + c_r/t_r$$

No primeiro passo esta consulta é lexicalmente ajustada. O objetivo deste ajuste é associar às expressões da consulta, descritores com aproximadamente o mesmo (possível) significado. Isto é feito por comparação de radicais, pois o radical é o portador do significado de uma palavra. Através da composição da consulta C com a relação difusa EE consegue-se esta consulta lexicalmente ajustada: $C_l = EE_C \circ C$,

onde EE_C é a relação entre expressões, *EE*, restrita às expressões da consulta e aos descritores. Neste modelo todas as composições são consideradas max-min composições, pois segundo Klir e Yuan [KLI 95], utiliza-se para a recuperação de documentos geralmente a composição *max-min*, portanto

$$C_I(x_i) = \max_{C(t)>0} \min\left(EE_C(x_i,t), C(t)\right), \text{para todo } x_i \in X.$$

Como no modelo de Miyamoto, aumenta-se esta consulta semanticamente através do thesaurus difuso: $C_{I,s} = T \circ C_I$:

$$C_{I,s}\left(x_j\right) = \max_{x \in X} \min\left(T\left(x_j, x\right), C_I(x)\right), \text{para todo } x_j \in X.$$

A partir daqui consegue-se a resposta como no modelo de Miyamoto, determinando primeiro a relevância dos documentos para os termos da consulta

$$R(d_j) = \left(V \circ C_{I,s}\right)(d_j) = \max_{x \in X} \min\left(V(d_j, x), C_{I,s}(x)\right), \text{para todo } d_j \in D,$$

e depois, como filtro, o corte-α de *D*, onde α, $0 \leq \alpha \leq 1$, é um parâmetro especificado pelo usuário,

$$^{\alpha}R = \{d \in D \mid R(d) \geq \alpha\}.$$

A resposta da consulta *C* é uma listagem dos documentos de $^{\alpha}R$ em ordem decrescente dos valores de pertinência em *R*.

Exemplo 4.2

Para mostrar o processo de recuperação de documentos considera-se um exemplo simples envolvendo somente palavras, e não expressões, com 6 termos indexados. Para simplificar o exemplo, supõe-se que o sistema tenha certeza 1 sobre a divisão dos termos indexados. Seja

x1 = lógica = (lógic,a) *x2 = complexidade = (complex,idade)*

x3 = equação = (equaç,ão) *x4 = compilação = (compil,ação)*

x5 = composição = (comp,osição) *x6 = implicação = (implic,ação)*

A consulta consiste de 3 termos de consulta:

t_1 t_2 t_3

C = 0.7/equações + 0.9/compilador + 0.8/competicão

A aplicação do algoritmo de reconhecimento de radicais e sufixos poderia ter como resultado:

tabela(equações) = 0.7/(equ,ações) + 0.8/(equaç,ões)
tabela(compilador) = 0.5/(comp,ilador) + 0.8/(compil,ador) + 0.7/(compilad,or)
tabela(competição) = 0.3/(comp,etição) + 0.8/(compet,ição) + 0.7/(competiç,ão)

onde tabela(p) é a tabela associada à palavra p considerada em 4.1.

A relação difusa entre os termos indexados e os termos da consulta determina-se como:

$$PP_C = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{pmatrix} 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.3 & 0 \end{pmatrix} \end{array} \begin{matrix} t_1 \\ t_2 \\ t_3 \end{matrix},$$

onde as linhas referem-se aos termos de consulta t_1, t_2 e t_3 e as colunas aos termos indexados $x_1, x_2, \ldots x_6$. Então a consulta lexicalmente ajustada é determinada como:

$$C_l = EE_C \circ C = PP_C \circ C = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{pmatrix} 0 & 0 & 0.7 & 0.8 & 0.3 & 0 \end{pmatrix} \end{array},$$

Seja o thesaurus difuso, T, determinado como:

$$T^{-1} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{pmatrix} 0.5 & 0.4 & 1 & 0.2 & 0.3 & 0.1 \\ 0.2 & 0.3 & 0.2 & 1 & 0.3 & 0.1 \\ 0.3 & 0.7 & 0.3 & 0.3 & 1 & 0.3 \end{pmatrix} \end{array} \begin{matrix} x_3 \\ x_4 \\ x_5 \end{matrix},$$

onde somente as linhas de T com $C_l(x_i) > 0$ são anotadas, i.e. as linhas associadas a x_3, x_4, x_5.

Determina-se como consulta semanticamente aumentada:

$$C_{l,s} = T \circ C_l = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{pmatrix} 0.5 & 0.4 & 0.7 & 0.8 & 0.3 & 0.3 \end{pmatrix} \end{array}$$

Seja a relação de relevância dada pela matriz:

$$V^{-1} = \begin{array}{c} \begin{matrix} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \end{matrix} \\ \begin{pmatrix} 0.2 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0.3 & 0 & 0.4 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.8 & 0 & 0.4 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.7 & 0.5 \\ 1 & 0 & 0.5 & 0 & 0 & 0.6 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.2 & 0 & 1 & 0 & 0 & 0.5 \end{pmatrix} \end{array} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{matrix},$$

onde os d_j são os documentos relevantes para a consulta. Então pode-se determinar:

$$\begin{array}{c} \underline{d_1 \quad d_2 \quad d_3 \quad d_4 \quad d_5 \quad d_6 \quad d_7 \quad d_8 \quad d_9 \quad d_{10}} \\ R = V \circ C_{I,S} = \quad (0.4 \quad 0.8 \quad 0.7 \quad 0.3 \quad 0.4 \quad 0.4 \quad 0.7 \quad 0.8 \quad 0.7 \quad 0.5) \end{array}$$

e considerando como filtro um corte-α de 0.7, tem-se um conjunto de cinco elementos:

$$^{0.7}R = \{d_2, d_3, d_7, d_8, d_9\}$$

Colocando estes documentos em ordem decrescente dos valores de pertinência tem-se como resposta:

$$d_2, d_8, d_3, d_7, d_9.$$

4.3.1 Construção do thesaurus difuso

A construção do thesaurus difuso é baseada na relação de proximidade gerada a partir das freqüências de co-ocorrências, orientada nos modelos de Miyamoto, Miyake e Nakayama [MIY 83].

Seja I: $X \times D \rightarrow \{0,1\}$ um índice que associa a cada documento os termos indexados, isto é $I(x_i,d_j) = 1$ significa que x_i é descritor do documento d_j, e $I(x_i,d_j) = 0$ significa que x_i não é descritor do documento d_j. Aplicando um ajuste lexical, como foi feito no ajuste lexical das consultas, define-se uma relação difusa H: $X \times D \rightarrow [0,1]$, um 'índice difuso', das associações entre documentos e termos indexados.

$$H(x_i,d_j) := \max_{x \in X} min\left(EE_X(x_i,x), I(x,d_j)\right),$$

onde EE_X: $X \times X \rightarrow [0,1]$ é a relação entre expressões, EE, restrita aos descritores.

A partir desta relação define-se o thesaurus, T: $X \times X \rightarrow [0,1]$, considerando freqüências de co-ocorrências de associações entre descritores e documentos. Portanto,

$$T(x_i,x_j) := \frac{\sum_{s=1}^{m} min\left(H(x_i,d_s), H(x_j,d_s)\right)}{\sum_{s=1}^{m} max\left(H(x_i,d_s), H(x_j,d_s)\right)}$$

Como para todo descritor x existe no mínimo um documento d com $I(x,d)=1$, tem-se $H(x,d)=1$ para este documento, pois a relação EE_X é reflexiva. Portanto, o denominador de T é diferente de 0 para cada dupla de termos indexados e, conseqüentemente, T é bem definido.

Exemplo 4.3

Seja a relação binária, $I: X \times D \rightarrow \{0,1\}$, que associa a cada documento os descritores, determinada como

$$I^{-1} = \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \left(\begin{array}{cccccc} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array} \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \end{array}$$

e seja a relação difusa, $EE_X : X \times X \rightarrow [0,1]$, entre os termos indexados dada por

$$EE_X = \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \left(\begin{array}{cccccc} 1 & 0 & 0 & 0.8 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.6 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.8 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0.6 & 0 & 0 & 0 & 1 \end{array}\right) \end{array} \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array}$$

Então determina-se a relação difusa, $H: X \times D \rightarrow [0,1]$, de associação entre os termos indexados e os documentos como

$$H^{-1} = (EE_X \circ I)^{-1} = \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \left(\begin{array}{cccccc} 1 & 1 & 0 & 0.8 & 0 & 0.6 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.8 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0.6 & 0 & 0.8 & 0 & 1 \end{array}\right) \end{array} \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \end{array}$$

Os valores de pertinência do pseudo-thesaurus difuso calculam-se, por exemplo, como:

$$T_{1,2} = \frac{min(1;1) + min(0;0) + min(0.8;0) + min(1;0.6)}{max(1;1) + max(0;0) + max(0.8;0) + max(1;0.6)} = \frac{1.6}{2.8} \approx 0.6$$

$$T_{1,4} = \frac{min(1;0.8) + min(0;0) + min(0.8;1) + min(1;0.8)}{max(1;0.8) + max(0;0) + max(0.8;1) + max(1;0.8)} = \frac{2.4}{3} \approx 0.8$$

Considerando todas as combinações, arredondando-se os resultados da divisão, tem-se como pseudo-thesaurus $T: X \times X \rightarrow [0,1]$,

$$T = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{pmatrix} 1 & 0.6 & 0 & 0.8 & 0.3 & 0.6 \\ 0.6 & 1 & 0 & 0.5 & 0 & 0.6 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.8 & 0.5 & 0 & 1 & 0.4 & 0.8 \\ 0.3 & 0 & 0 & 0.4 & 1 & 0 \\ 0.6 & 0.6 & 0 & 0.8 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{matrix}$$

4.3.2 Construção da relação de relevância dos termos indexados para os documentos

A construção da relação de relevância dos termos indexados (descritores) para os documentos é estabelecida de forma semelhante a proposta por Miyamoto e Nakayama em [MIY 86], substituindo o índice não difuso $I: X \times D \rightarrow \{0,1\}$ pela relação difusa $H: X \times D \rightarrow [0,1]$.

A relação difusa de relevância dos termos indexados para os documentos é definida a partir do thesaurus difuso e a relação difusa H, que associa a cada documento os termos indexados. O grau de relevância de um termo indexado para um documento pode ser considerado o máximo dos valores de pertinência no thesaurus difuso entre este termo em consideração e os termos associados ao documento, i.e. $V: X \times D \rightarrow [0,1]$, com

$$V(x_i, d_j) := \max_{x \in X} \min\left(T(x_i, x), H(x, d_j)\right)$$

Exemplo 4.4

Considerando a relação difusa H e o thesaurus T do exemplo 4.3 determina-se como relação de relevância, $V: X \times D \rightarrow [0,1]$, dos termos indexados para os documentos

$$V^{-1} = (T \circ H)^{-1} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 0.8 & 0.4 & 0.8 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.8 & 0.6 & 0 & 1 & 1 & 0.8 \\ 1 & 0.6 & 0 & 0.8 & 0.4 & 1 \end{pmatrix} \end{array} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{matrix}$$

4.4 Um exemplo completo

Para demonstrar o procedimento completo, considera-se um exemplo com somente 4 documentos, d_1, d_2, d_3, d_4. A estes 4 documentos estão associadas as seguintes expressões:

$d_1 \leftrightarrow x_1$ = *'teoria difusa'*, x_2 = *'recuperar informações'*,

$d_2 \leftrightarrow x_3$ = *'relações difusas'*

$d_3 \leftrightarrow x_4$ = *'teoria de informação'*, x_5 = *'documento'*

$d_4 \leftrightarrow x_1$ = *'teoria difusa'*, x_6 = *'recuperação de documentos'*

Então determina-se como relação binária entre documentos e descritores exatamente a relação binária, $I: X \times D \rightarrow \{0,1\}$, do exemplo 4.3.

A aplicação do algoritmo de reconhecimento de radicais e sufixos poderia ter como resultado:

tabela(teoria) = 0.8/(teor,ia) + 0.7/(teori,a)

tabela(difusa) = 0.8/(dif,usa) + 0.9/(difus,a)

tabela(recuperar) = 0.8/(recup,erar) + 0.9/(recuper,ar)

tabela(informações) = 0.7/(inform,ações) + 0.8/(informaç,ões) + 0.5/(informaçõ,es)

tabela(relações) = 0.5/(rel,ações) + 0.8/(relaç,ões) + 0.7/(relaçõ,es)

tabela(difusas) = 0.6/(dif,usas) + 0.8/(difus,as)

tabela(informação) = 0.6/(inform,ação) + 0.7/(informaç,ão) + 0.4/(informaçã,o)

tabela(documento) = 0.6/(doc,umento) + 0.7/(docum,ento) + 0.5/(document,o)

tabela(recuperação) = 0.5/(recup,eração) + 0.8/(recuper,ação) + 0.7/(recuperaç,ão)

tabela(documentos) = 0.3/(doc,umentos) + 0.7/(docum,entos) + 0.8/(document,os)

Considerando as seguintes atribuições aos termos dos descritores

$\{ p_1, \ldots, p_{10} \}$ = {*'teoria'*, *'difusa'*, *'recuperar'*, *'informações'*, *'relações'*, *'difusas'*, *'informação'*, *'documento'*, *'recuperação'*, *'documentos'*}

determina-se como relação entre as palavras abertas dos descritores

$$PP_X = \begin{array}{c} \begin{array}{cccccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0.7 \\ 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{l} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \\ p_9 \\ p_{10} \end{array}$$

Portanto, determina-se como relação, $EE_X : X \times X \rightarrow [0,1]$, entre os descritores exatamente a relação considerada no exemplo 4.3, considerando as penalidades do exemplo 4.1. Conseqüentemente, determina-se o thesaurus difuso, $T: X \times X \rightarrow [0,1]$, como no exemplo 4.3 e a relação de relevância, $V: X \times D \rightarrow [0,1]$, dos termos indexados para os documentos como no exemplo 4.4.

Para simplificar o exemplo, considera-se uma consulta que envolve somente uma expresssão

$$\begin{array}{cc} t_1 & t_2 \end{array}$$
$$e = \text{'recuperando informação'}; \qquad C = 1/e$$

O resultado do algoritmo de reconhecimento de radicais e sufixos poderia ser:

tabela(recuperando) = 0.2/(rec,uperando) + 0.4/(recup,erando) + 0.8/(recuper/ando)

tabela(informação) = 0.6/(inform,ação) + 0.7/(informaç,ão) + 0.4/(informaçã,o)

Então, determina-se como relação, PP_C, entre os termos dos descritores e os termos da consulta:

$$PP_C = \begin{array}{c} \begin{array}{cccccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} \end{array} \\ \begin{pmatrix} 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.7 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{array}{l} t_1 \\ t_2 \end{array}$$

e como relação entre a expressão da consulta e os descritores

$$\begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array}$$
$$EE_C = \begin{pmatrix} 0 & 0.7 & 0 & 0 & 0 & 0.6 \end{pmatrix}$$

A consulta lexicalmente ajustada determina-se como

$$C_l = EE_C \circ C \quad = EE_C$$

porque a consulta C tem somente uma expressão com grau de pertinência 1.

Como no sistema de Miyamoto, extende-se a consulta aplicando o thesaurus difuso T:

$$\begin{array}{lll} & \underline{x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6} \\ C_{l,s} = T \circ C_l & = & (0.6 \quad 0.7 \quad 0 \quad 0.6 \quad 0 \quad 0.6) \end{array}$$

Aplicando a relação, V, de relevância dos termos indexados para os documentos determina-se:

$$\begin{array}{lll} & \underline{d_1 \quad d_2 \quad d_3 \quad d_4} \\ R = V \circ C_{l,s} & = & (0.7 \quad 0 \quad 0.6 \quad 0.6) \end{array}$$

Considerando um corte-α de 0.6 como filtro, tem-se como resultado:

$$d_1, d_3, d_4.$$

Analisando este resultado pode-se observar que:

- os documentos d_1 e d_4 são retornados embora nenhuma palavra dos descritores coincida com as palavras da consulta;
- o documento d_2 não é retornado, embora de fato exista uma ligação entre relações difusas e a recuperação de informação, mas não exista nada que poderia associar estes descritores por co-ocorrência;

4.5 Considerações finais do capítulo

Comparando este modelo com o modelo genérico de Murai, Miyakoshi e Shimbo [MUR 89] pode-se observar que neste modelo a resposta é determinada pela aplicação de três relações difusas. A relação EE_C é somente determinada depois de realizada a consulta. Portanto, pode-se considerar a aplicação de EE_C como uma transformação da consulta em uma representação a que se pode aplicar a função de recuperação, i.e. em termos de Murai, Miyakoshi e Shimbo [MUR 89]

$$\chi(C) = EE_C \circ C$$

$$\rho(C', \bullet) = V \circ T \circ C'$$

$$R = \{(d,s) \in D \times [0,1] \mid \rho(\chi(C),d) = s\}$$

A tabela dos conectores considerados neste modelo (figura 4.3), apresenta muito mais ambigüidades do que no modelo de Bruza e van der Weide (figura 3.2). Por exemplo, a preposição 'de' aparece em todas as relações consideradas. Isto poderia ser devido ao fato de que, talvez por razões históricas, o papel das preposições na língua inglesa é mais restrito do que na lingua portuguesa. Considerou-se a possibilidade de caminhos percorrendo arcos em direção contrária,

porque na língua portuguesa a posição de uma palavra em uma expressão não modifica, necessariamente, o significado da expressão. Atribui-se penalidades neste caso, pois a troca de posição de palavras pode levar a modificações do significado como, por exemplo, em: filho meu - meu filho; grande homem - homem grande. Entendeu-se como generalização (especialização) de uma expressão a inclusão (remoção) de uma dupla conector, termo. Como generalização e especialização de uma expressão não tem significado igual ao significado da expressão, atribui-se também uma penalidade para palavras abertas não consideradas na comparação.

A relação entre palavras *PP* é semelhante a uma das medidas considerada por Farreny e Prade [FAR 86] para estimar similaridades dos significados entre valores de variáveis lingüísticas. Os autores interpretam que esta medida estima a possibilidade de ter um significado similar.

5. AVALIAÇÃO DO MODELO

A avaliação de um modelo de recuperação de documentos é importante, porque a decisão de operar um sistema segundo um determinado modelo depende, em última instância, desta questão. Segundo Salton [SAL 75] e Salton e McGill [SAL 83], a efetividade e a eficiência são evidenciadas na avaliação de um sistema de recuperação de informação;

- a efetividade é a capacidade de fornecer as informações que o usuário precisa;
- a eficiência é uma medida do custo e/ou tempo necessário para efetuar determinadas tarefas.

A efetividade e a eficiência podem ainda ser examinadas do ponto de vista dos usuários ou do ponto de vista dos operadores do sistema. Segundo Salton e McGill [SAL 83], os critérios de efetividade dos usuários e dos operadores são semelhantes, pois pode-se supor que um sistema de informação existe para satisfazer as necessidades de usuários. A respeito do custo e benefício do sistema, os interesses dos operadores e dos usuários são também análogos até um determinado ponto.

Entre os vários critérios de avaliação de um sistema de informação do ponto de vista dos usuários, Salton [SAL 75], Salton e McGill [SAL 83] e Lancaster [LAN 87], apontam seis critérios como cruciais :

- a evocação (recall), i.e. a capacidade de extrair aqueles documentos que podem ser considerados relevantes;
- a precisão, i.e. a capacidade de evitar a recuperação de documentos que não sejam relevantes;
- o esforço necessário para formular as consultas, conduzir a pesquisa e examinar o resultado;
- o tempo de resposta, i.e. o intervalo de tempo entre a entrega da consulta e a apresentação do resultado;
- a forma de apresentação do resultado, o que influencia a habilidade do usuário em utilizar as informações recuperadas;
- a completude, i.e. a extensão em que o sistema inclui todos os itens relevantes.

Segundo Salton [SAL 75] e Salton e McGill [SAL 83], o esforço exigido do usuário e o tempo de execução podem ser medidos facilmente. O esforço exigido do usuário pode ser medido como o tempo necessário para formular as consultas, interagir com o sistema e examinar o resultado. O tempo de resposta pode ser medido diretamente como o tempo de processamento do sistema, ou seja em sistemas computacionais como complexidade do algoritmo. O critério da completude e a forma de apresentação dos resultados dependem da base de dados, portanto não serão considerados aqui. Para os

autores, a determinação das medidas de evocação e precisão apresentam os maiores problemas.

5.1. Evocação e precisão

O primeiro problema na determinação da evocação e da precisão é a interpretação da palavra 'relevante'. Segundo Salton e McGill [SAL 83], pelo menos duas definições da relevância são possíveis:

- uma, objetiva, que considera a relevância uma característica lógica entre consulta e documento. Ela pode ser medida considerando se o documento trata ou não do assunto da consulta;
- uma, mais subjetiva, que não considera somente o conteúdo do documento mas também o conhecimento do usuário no momento da pesquisa.

É mais fácil determinar a relevância através da definição objetiva. Mas mesmo considerando esta definição, pode-se ter desacordo entre vários avaliadores sobre o grau de relevância de um documento para uma determinada consulta. Além disso, as necessidades de informação podem variar entre os usuários. Um usuário pode querer uma evocação alta, i.e. a recuperação de quase tudo que poderia ser de interesse, um outro pode preferir uma precisão alta, i.e. a rejeição de tudo que poderia ser inútil.

Na prática, utiliza-se um critério externo para determinar a relevância, como, por exemplo, a opinião de especialistas. Assim, define-se

- evocação como a proporção entre os documentos relevantes e recuperados e os documentos relevantes na base de dados, i. e.

$$evocação = \frac{número\ de\ itens\ recuperados\ e\ relevantes}{número\ de\ itens\ relevantes\ na\ base\ de\ documentos}$$

- precisão como a proporção entre os documentos relevantes e recuperados e os documentos recuperados, i.e.

$$precisão = \frac{número\ de\ itens\ recuperados\ e\ relevantes}{número\ total\ de\ itens\ recuperados}$$

Um sistema é considerado bom, se ele apresenta altos valores de evocação e altos valores de precisão. Em sistemas reais observa-se que buscas genéricas tendem a recuperar muitos itens, i.e. evocação alta e precisão baixa, enquanto buscas específicas recuperam poucos itens, i.e. evocação baixa e precisão alta. Para determinar a relação entre evocação e precisão devia-se, por exemplo, determinar várias consultas com valores de evocação cobrindo o intervalo de [0,1], e determinar a precisão destas consultas. Como isto é geralmente impossível, determina-se gráficos de evocação/precisão para cada consulta considerando a precisão nos diferentes níveis de evocação.

5.1.1 Gráficos de evocação/precisão de uma consulta

Para a determinação dos gráficos de evocação/precisão de uma consulta determina-se a evocação e a precisão percorrendo a lista das respostas, considerando sempre somente os documentos recuperados na sub-lista percorrida. A figura 5.1 mostra evocação e precisão para uma única consulta com 14 documentos recuperados.

n	nº do documento	x = relevante	evocação	precisão
1	588	x	0.2	1.0
2	589	x	0.4	1.0
3	576		0.4	0.67
4	590	x	0.6	0.75
5	986		0.6	0.60
6	592	x	0.8	0.67
7	984		0.8	0.57
8	988		0.8	0.50
9	578		0.8	0.44
10	985		0.8	0.40
11	103		0.8	0.36
12	591		0.8	0.33
13	772	x	1.0	0.38
14	990		1.0	0.36

Figura 5.1: Evocação e precisão para uma única consulta depois de n documentos recuperados (figura 5.2 (a) em [SAL 83]).

O exemplo da figura 5.1 considera 5 documentos relevantes. Os valores de evocação e precisão são determinados considerando somente os documentos anteriores na lista da resposta. Por exemplo, para a determinação da evocação e precisão após a recuperação do documento 576, considera-se dois documentos relevantes e um documento não relevante. Portanto tem-se: evocação = 2 / 5 = 0.4 e precisão = 2 / 3 = 0.67.

Representando em um gráfico os pares evocação/precisão determinados para uma consulta constata-se que existem diversos valores de precisão para um único valor de evocação. Portanto considera-se interpolações. Existem diversas possibilidades de interpolar estas curvas; uma forma muito simples, indicada por Salton [SAL 75] e Salton e McGill [SAL 83], é desenhar a partir de cada pico uma linha horizontal na direção esquerda até encontrar a curva original. A figura 5.2 mostra o gráfico dos pares evocação/precisão do exemplo da figura 5.1 e a interpolação deste gráfico. A interpolação tem, por exemplo, uma precisão de 0.75 para valores de evocação no intervalo *(0.4;0.6]*, e uma precisão de 1 para a evocação de 0.4. Segundo Salton e McGill [SAL 83], este tipo de interpolação representa o melhor desempenho que um usuário pode esperar.

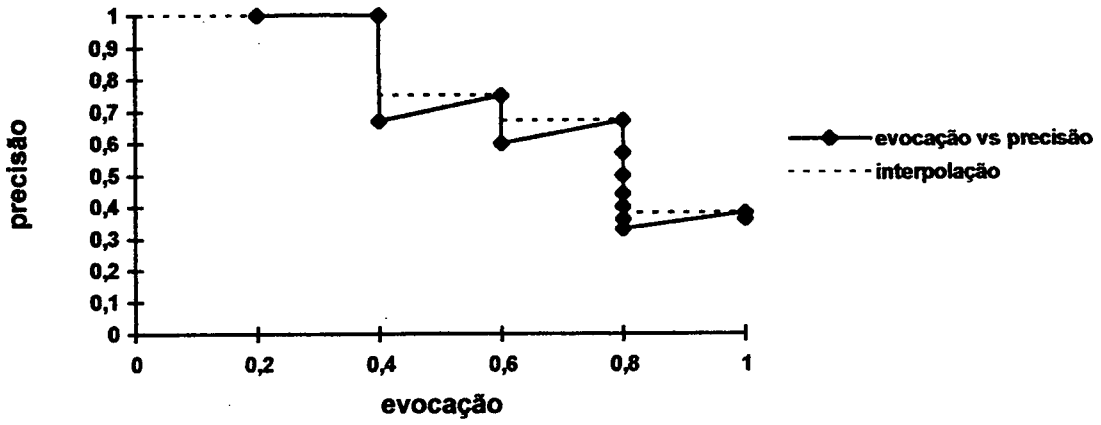


Figura 5.2: Gráfico dos pares evocação/precisão do exemplo da figura 5.1 e sua interpolação.

Ao contrário das medidas de evocação e precisão, a sequência das respostas é importante nos gráficos de evocação/precisão. Trocando, por exemplo, o documento relevante 589 na posição 2 com o documento não-relevante 985 na posição 10, no exemplo da figura 5.1, determina-se o gráfico de evocação/precisão e sua interpolação como mostrado na figura 5.3.

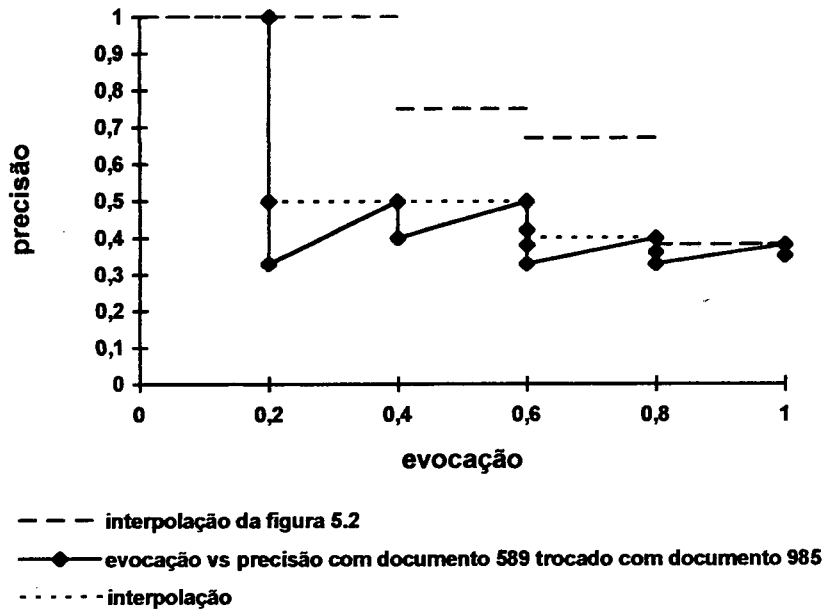


Figura 5.3 Gráfico de evocação/precisão para o exemplo da figura 5.1 trocando as posições dos documentos 589 e 985, sua interpolação e a interpolação do gráfico de evocação/precisão para o exemplo da figura 5.1 (interpolação da figura 5.2).

A figura 5.3 mostra que a curva de evocação/precisão com o documento relevante 589 na posição 2 fica bem acima da curva de evocação/precisão com este documento somente na posição 10 da resposta, indicando um melhor desempenho.

5.1.2 Médias de evocação/precisão

Evocação e precisão determinadas a partir de uma única consulta não consideram vários parâmetros importantes como, por exemplo, o tamanho do conjunto de documentos recuperados. Por isto determina-se médias de evocação e precisão, considerando um número fixo de consultas.

Considerando k consultas define-se para $1 \leq i \leq k$

a_i = número de itens recuperados e relevantes para a consulta i;

b_i = número de itens recuperados e não-relevantes para a consulta i;

c_i = número de itens relevantes, mas não recuperados para a consulta i;

- as médias usuário-orientadas de evocação e de precisão são definidas como

$$evocação_{uo} = \frac{1}{k}\sum_{i=1}^{k}\frac{a_i}{a_i + c_i}$$

$$precisão_{uo} = \frac{1}{k}\sum_{i=1}^{k}\frac{a_i}{a_i + b_i}$$

Essas médias refletem o desempenho que um usuário pode esperar, em média, do sistema de recuperação.

- as médias sistema-orientadas de evocação e de precisão são obtidas considerando o número total dos itens recuperados nas k consultas, imaginando uma consulta hipotética, composta, i.e.

$$evocação_{so} = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k}(a_i + c_i)}$$

$$precisão_{so} = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k}(a_i + b_i)}$$

As médias usuário-orientadas consideram todas as consultas com importância igual, enquanto as médias sistema-orientadas dependem mais das consultas com muitos documentos relevantes do que das consultas com poucos documentos relevantes.

Além disso pode-se determinar os gráficos de evocação/precisão em média para as k consultas. Isto pode ser feito determinando médias de precisão em valores fixos de evocação - por exemplo 0, 0.1, 0.2, ..., 1.0 - como médias das interpolações dos gráficos de evocação/precisão nestes valores fixos de evocação. A figura 5.4 mostra um gráfico típico de evocação/precisão em média como ele é apresentado por diversos autores, por exemplo [SAL 75], [LOP 96].

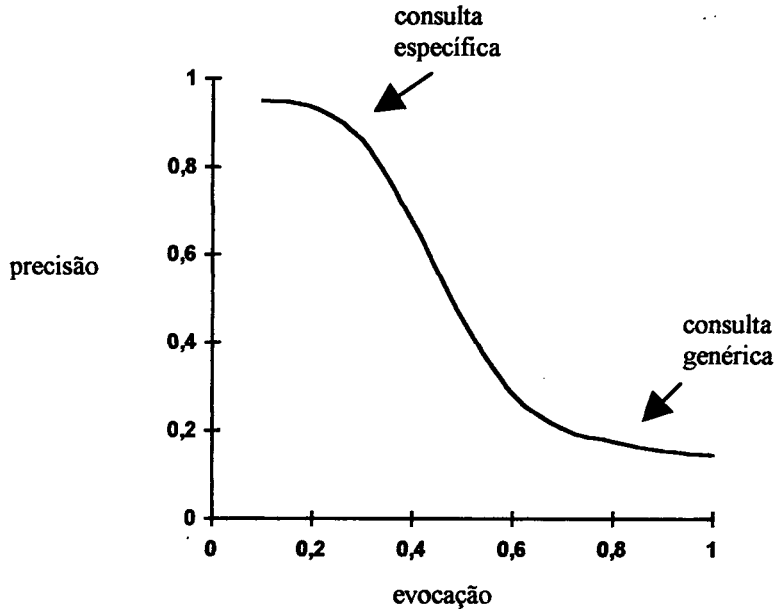


Figura 5.4: Gráfico típico de evocação/precisão em média.

5.1.3 Outras medidas de efetividade

Às vezes é impossível calcular a evocação e a precisão. Por exemplo, a evocação não é definida se não existe um documento relevante para uma consulta. A precisão não é definida se nenhum documento é recuperado para uma consulta. Por isso, dois pârametros adicionais foram introduzidos para medir a efetividade de um sistema de recuperação: o fallout e a generalidade.

$$fallout = \frac{número\ de\ itens\ não\text{-}relevantes\ recuperados}{número\ de\ itens\ não\text{-}relevantes\ na\ base\ de\ documentos}$$

$$generalidade = \frac{número\ de\ itens\ relevantes}{número\ de\ itens\ na\ base\ de\ documentos}$$

O fallout mede o comportamento dos documentos não relevantes da mesma maneira como a evocação mede o comportamento dos documentos relevantes. Para medir a efetividade de um sistema de recuperação pode-se, por exemplo, substituir a precisão pelo fallout e determinar médias da evocação e fallout, bem como gráficos de evocação e fallout. Um sistema de recuperação de documentos efetivo apresenta uma evocação máxima com um fallout mínimo. Existe uma relação entre os quatro parâmetros - evocação, precisão, fallout e generalidade - que permite determinar sempre o quarto parâmetro através dos outros três, por exemplo:

$$\text{precisão} = \frac{\text{evocação} \cdot \text{generalidade}}{(\text{evocação} \cdot \text{generalidade}) + \text{fallout}(1 - \text{generalidade})}$$

A literatura considera diversas outras medidas de efetividade de um sistema de informação, substituindo a evocação e precisão, como, por exemplo, a seletividade e a especificidade [SAL 75] e [SAL 83], medidas que combinam evocação e precisão [SAL

83] e [LOP 96] e probabilidade de relevância e expectância de precisão [RAG 89]. Estas medidas não são tratadas aqui, porque não foram utilizadas na avaliação do modelo proposto.

5.1.4. Comparação de modelos

Segundo Salton [SAL 75], as medidas de evocação e precisão dependem, até um determinado ponto, do tamanho da coleção de documentos e da média dos itens relevantes nas consultas consideradas. Por isso deve-se considerar na comparação de dois sistemas a mesma coleção de documentos e o mesmo ou equivalente conjunto de consultas . Além disso o conjunto de consultas considerado deve ser uma mistura realística das consultas aplicadas em situação de operação real.

Segundo Salton [SAL 75] e Salton e McGill [SAL 83], a significância das diferenças de dois modelos pode ser dificilmente avaliada através dos gráficos de evocação e precisão. Por isso utiliza-se testes estatísticos de significância baseados na comparação de pares de valores, como, por exemplo, o teste t , o teste dos sinais ou o teste de Wilcoxon.

Salton e McGill [Sal 83] afirmam que os pares de valores comparados podem ser valores de precisão em diferentes níveis de evocação ou o número de consultas favorecendo um ou outro método, i.e. o número de consultas com valores de precisão de um modelo maior do que do outro modelo nos diferentes níveis de evocação.

O teste t supõe que as diferenças dos pares comparados seguem a distribuição normal. Essas diferenças muitas vezes não satisfazem a condição da distribuição normal em sistemas de recuperação, e os autores afirmam ser preferível utilizar ou o teste do sinal ou o teste de Wilcoxon [SAL 75;83].

5.2. Comparação entre o modelo de Miyamoto e o modelo proposto

Para a comparação entre o modelo de Miyamoto e o modelo proposto, foi feita uma implementação em *Smalltalk/V for Windows* [DIG 92], seguindo o paradigma de programação orientada a objetos. A implementação do modelo proposto foi chamada *Radix*. Para os testes foi estabelecida uma base de 36 documentos (anexo 2), um vocabulário controlado com 77 expressões (anexo 3) e um dicionário de palavras da classe aberta contendo 158 palavras (anexo 4).

5.2.1 Implementação

A implementação dos dois sistemas foi feita através de 14 classes. A figura 5.3 mostra as 14 classes implementadas e a relação hierárquica entre as classes.

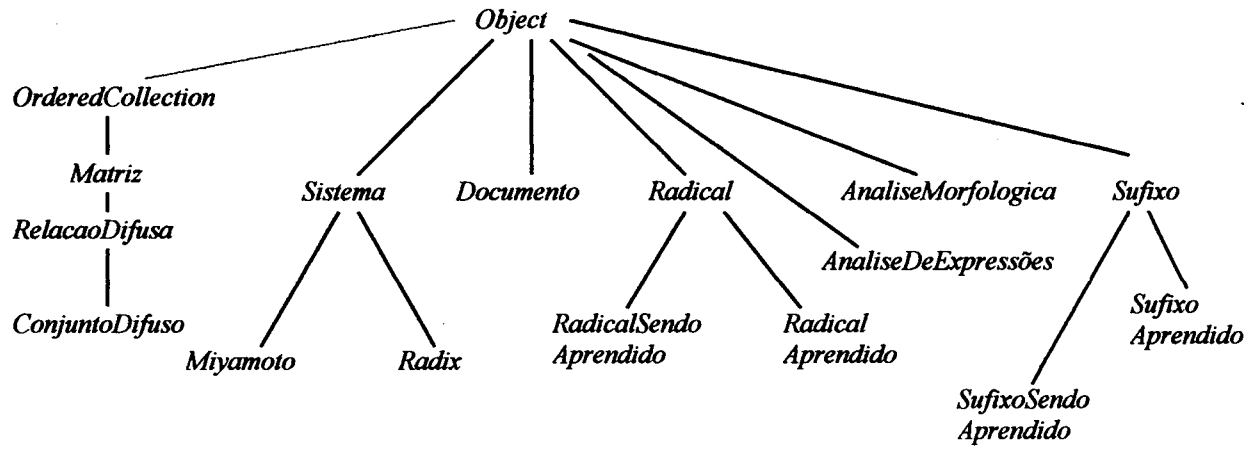


Figura 5.5: Relação hieráquica entre as classes implementadas considerando também as classes *Object* e *OrderedCollection* do Smalltalk. A linha tracejada indica a existência de outras classes entre a classe *Object* e a classe *OrderedCollection*.

A implementação considera uma relação difusa, uma matriz em que os elementos representam os valores de pertinência. Um conjunto difuso é interpretado como um caso especial de uma relação difusa de uma linha só. A implementação das classes *Miyamoto* e *Radix* é direta considerando as equações para a determinação dos índices, thesauri, relações de relevância e as respostas apresentadas nos capítulos 3 e 4. A implementação do modelo de Miyamoto considera a relação de proximidade s_1 do capítulo 3.2.1 para a determinação do thesaurus difuso. A implementação do sistema proposto considera penalidade de 0.1 para o percurso de um arco em direção errada ou o percurso de um arco em direção correta, mas o conector não pode representar a mesma relação, e a penalidade de 0.2 para cada palavra aberta ainda não considerada.

O sistema de reconhecimento de radicais e sufixos é basicamente o sistema apresentado em [STO 96] (anexo 1). Os objetos básicos do sistema são as palavras com seus radicais e sufixos, bem como dicionários que permitem acesso a informações sobre estes objetos.

5.2.1.1 Tratamento das certezas

Quando o sistema encontra uma palavra ainda não conhecida, ele, primeiro, cria uma tabela contendo todos os pares radical/sufixo. Por exemplo, a palavra 'difuso' seria dividida na lista de pares representada na figura 5.4.

Cada novo radical e novo sufixo recebe primeiramente uma certeza inicial:

$$certeza\_inicial_{radical} = certeza\_inicial_{sufixo} = ½$$

As certezas de radical e sufixo são reforçadas no caso de novas ocorrências de radical ou sufixo em uma análise segundo as fórmulas:

$$certeza_{radical} := certeza_{radical} + (1 - certeza_{radical})\ certeza_{sufixo}/2$$

$$certeza_{sufixo} := certeza_{sufixo} + (1 - certeza_{sufixo})\ certeza_{radical}/2$$

Elas são enfraquecidas no caso da remoção de uma possível combinação com este radical ou sufixo do dicionário das palavras.

$$certeza_{radical} := certeza_{radical} - certeza_{radical}\ certeza_{sufixo}/2$$

$$certeza_{sufixo} := certeza_{sufixo} - certeza_{sufixo}\ certeza_{radical}/2$$

Para cada par radical/sufixo é calculado um grau de certeza, um valor entre *0* e *1*, baseado na combinação das certezas obtidas para o radical e o sufixo. A combinação das certezas do radical e do sufixo é considerada uma variação da conjunção difusa, a função de certeza mínima com redução, sendo as certezas dos pares radical/sufixo determinadas por:

$$certeza_{par} = min\ \{certeza_{radical} * \alpha_{|radical|},\ certeza_{sufixo} * \alpha_{|sufixo|}\}$$

onde $\alpha_x = 3/4 + x/16$ se $0 \leq x \leq 3$ e $\alpha_x = 1$ se $x > 3$, e $|radical|$ e $|sufixo|$ significam os tamanhos do radical e do sufixo, ou seja, o número de letras que formam cada radical ou sufixo. A figura 5.6 mostra um exemplo de divisão da palavra "difuso" em possíveis radicais e a determinação das certezas de pares radical/sufixo.

radical	$certeza_{radical}$	sufixo	$certeza_{sufixo}$	$certeza_{par}$
d	0.6	ifuso	0.0	0.0
di	0.2	fuso	0.1	0.1
dif	1.0	uso	0.3	0.3
difu	0.5	so	0.1	0.1
difus	0.8	o	1.0	0.8
difuso	0.3		1.0	0.3

Figura 5.6 : Exemplo de divisão da palavra 'difuso' em pares radical/sufixo; certezas de radicais ($certeza_{radical}$); certezas de sufixos ($certeza_{sufixo}$) e certezas de pares radical/sufixo ($certeza_{par}$).

O processo de aprendizagem consiste na eliminação de possíveis pares radical/sufixo da tabela associada á palavra no dicionário de palavras. Um par radical/sufixo será removido se ao mesmo tempo

$$0.7 \leq max\{certeza_{par(i)} \mid par(i) \in tabela\}.$$

e

$$certeza_{par} < 1/2\ max\{certeza_{par(i)} \mid par(i) \in tabela\}.$$ [1]

No caso de já existir uma entrada para a palavra encontrada, considera-se somente as combinações contidas na tabela associada à palavra.

[1] Esta forma de remoção de pares radical/sufixo é mais conservativa do que as consideradas em [STO 96], e foi escolhida devido a resultados de testes adicionais .

Os resultados apresentados em [STO 96] mostram que este algoritmo leva a valores altos de certeza para os pares radical/sufixo corretos (anexo 1).

5.2.2 Os testes

Para os testes criou-se uma base de 36 documentos (anexo 2). Cada documento tem associado dois conjuntos de descritores:

- um conjunto de palavras-chave ou descritores controlados: isto representa uma descrição do documento seguindo o vocabulário controlado (anexo 3);
- um conjunto de expressões ou descritores não-controlados: isto representa uma descrição do documento seguindo os 'keywords' dos autores do documento.

O vocabulário controlado de 77 expressões (anexo 3) é uma unificação do vocabulário utilizado pelos autores na descrição dos documentos. Por exemplo, o vocabulário controlado contém somente a palavra-chave 'banco de dados' no lugar em que os autores utilizaram tanto 'banco de dados' quanto 'base de dados'.

Nos testes foram consideradas 8 consultas (anexo 5). Estas consultas foram formuladas uma vez seguindo o vocabulário controlado, as chamadas consultas controladas, e outra vez variando as palavras das consultas, as chamadas consultas não-controladas. Para todas as consultas foram determinados conjuntos de documentos relevantes (anexo 6).

Além disso criou-se um dicionário de palavras da classe aberta contendo 158 palavras aprendidas, isto significa que este dicionário contém informações sobre a divisão correta destas palavras em radical e sufixo (anexo 4).

Para os testes considerou-se os sete seguintes casos diferentes:

1. Miyamoto: considera o modelo de Miyamoto utilizando descritores controlados; as consultas seguem o vocabulário controlado;
2. Radix com descritores controlados, consultas controladas e dicionário vazio: considera o modelo proposto utilizando os descritores controlados; as consultas seguem o vocabulário controlado; o analisador morfológico inicia com um dicionário vazio;
3. Radix com descritores controlados, consultas controladas e dicionário não-vazio: considera o modelo proposto utilizando os descritores controlados; as consultas seguem o vocabulário controlado; o analisador morfológico inicia com um dicionário de 158 palavras abertas aprendidas (anexo 4);
4. Radix com descritores controlados, consultas não-controladas e dicionário vazio: considera o modelo proposto utilizando os descritores controlados; as consultas não seguem o vocabulário controlado; o analisador morfológico inicia com um dicionário vazio;
5. Radix com descritores controlados, consultas não-controladas e dicionário não-vazio: considera o modelo proposto utilizando os descritores controlados; as consultas não

seguem o vocabulário controlado; o analisador morfológico inicia com um dicionário de 158 palavras abertas aprendidas (anexo 4);

6. Radix com descritores não-controlados, consultas não-controladas e dicionário vazio: considera o modelo proposto utilizando os descritores não-controlados; as consultas não seguem o vocabulário controlado; o analisador morfológico inicia com um dicionário vazio;
7. Radix com descritores não-controlados, consultas não-controladas e dicionário não-vazio: considera o modelo proposto utilizando os descritores não-controlados; as consultas não seguem o vocabulário controlado; o analisador morfológico inicia com um dicionário de 158 palavras abertas aprendidas (anexo 4).

Para todas estas $7 \cdot 8 = 56$ consultas foram determinados os gráficos de precisão nos níveis da evocação e os gráficos do fallout nos níveis da evocação. O anexo 7 apresenta exemplos de respostas a consultas e os gráficos de precisão nos níveis da evocação, bem como do fallout nos níveis de evocação. A partir daí determinou-se as médias (usuário-orientadas) de evocação, precisão e fallout nos níveis do threshold(α), nos dez casos de α = 0.1, 0.2, ... , 0.9, 1.0 (anexo 8). Quer dizer, considerou-se para a determinação da evocação, precisão e fallout somente os documentos com um valor de relevância determinado pelo sistema maior ou igual a α.

Esta forma de avaliar um sistema difuso de recuperação de documentos, considerando as médias (usuário-orientadas) nos níveis do threshold foi também utilizado por Shyi-Ming Chen e Jeng-Yih Wang [CHE 95]. No caso de consultas não-controladas existe o problema que o conjunto de respostas é, em alguns casos, vazio para valores altos do threshold α. Considerou-se nestes casos a precisão como não definida e determinou-se as médias de precisão somente considerando as consultas com um conjunto de respostas não-vazio. Isto leva a uma determinada distorção na comparação dos casos considerados. Por isso considerou-se preferível utilizar o fallout nos níveis do threshold para a avaliação dos casos com consultas não-controladas .

Para avaliar a seqüência das respostas determinou-se também nos sete casos as médias da precisão e do fallout nos níveis da evocação e os gráficos de evocação/precisão e evocação/fallout em média (anexo 8).

5.2.3 Os resultados

Os sete casos dos testes foram submetidos às três seguintes comparações:

- Miyamoto versus Radix com descritores controlados e consultas controladas: assim compara-se o thesaurus difuso determinado simplesmente pelas co-ocorrências dos descritores (Miyamoto, cap 3.2.1, relação s_1) com o thesaurus difuso determinado através do índice difuso (Radix, cap. 4.3.1).
- Radix com descritores controlados e consultas não-controladas versus Radix com descritores não-controlados e consultas não-controladas: serve para determinar a influência de um vocabulário controlado nos descritores caso não existam restrições nas consultas.

- Miyamoto versus Radix com descritores não-controlados e consultas não controladas: compara dois extremos, o modelo de Miyamoto que exige um vocabulário controlado e o modelo proposto sem restrições no vocabulário.

5.2.3.1 Miyamoto vs Radix com descritores controlados e consultas controladas

O gráfico apresentado na figura 5.7 mostra as médias de evocação e de precisão nos níveis do threshold para o modelo de Miyamoto e para o modelo proposto considerando descritores controlados e consultas controladas. Considerou-se aqui somente o caso em que o analisador inicia com as 158 palavras abertas aprendidas, pois a evocação e a precisão apresentaram pouca diferença entre os casos dicionário vazio e dicionário não-vazio (anexo 8).

Observa-se na figura 5.7 que os sistemas Radix e Miyamoto apresentaram, para valores do threshold α = 0.9 e 1.0, igual evocação e precisão, mas que o sistema Radix apresentou maior evocação e menor precisão do que o sistema Miyamoto para valores de α entre 0.1 e 0.8.

A figura 5.8 mostra os gráficos de evocação/precisão em média determinados para os casos em consideração. A figura mostra que o sistema Radix apresentou uma maior precisão nos níveis da evocação, para valores de evocação entre 0.2 e 0.4 e para valores de evocação entre 0.6 e 1.0. Para valores de evocação entre 0.1 e 0.2 e entre 0.4 e 0.6, os gráficos mostram precisão quase igual para os dois sistemas.

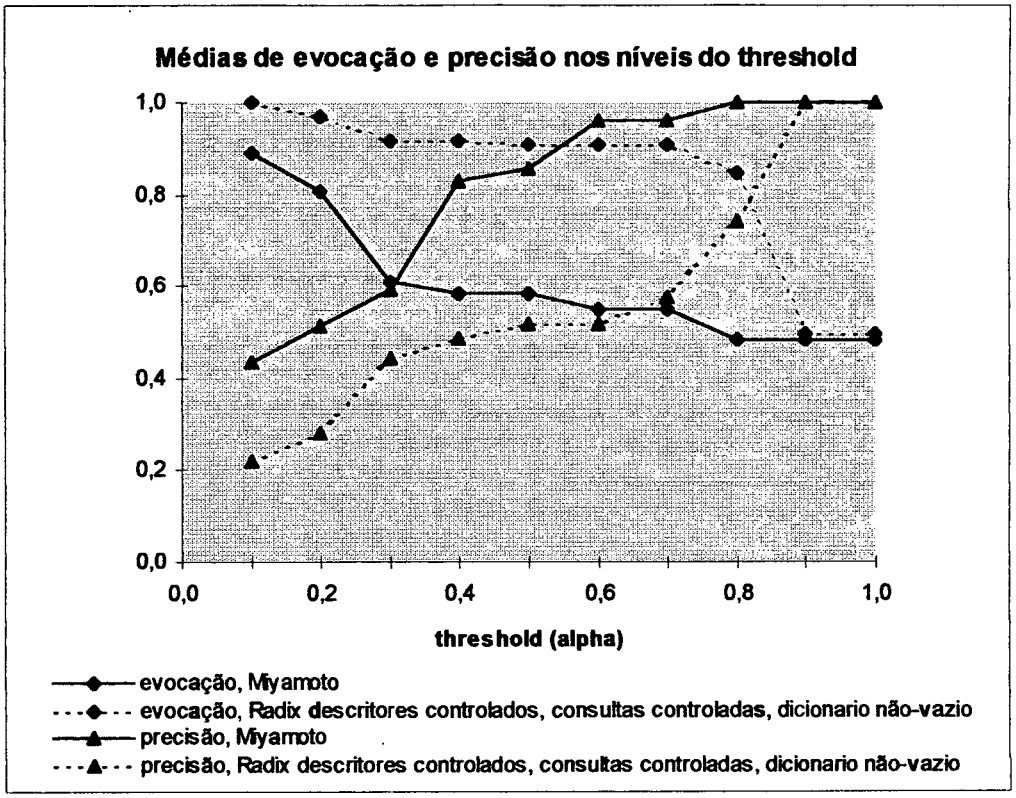


Figura 5.7: Médias de evocação e precisão nos níveis do threshold para o modelo de Miyamoto e para o sistema Radix considerando descritores controlados, consultas controladas e iniciando com um conjunto de palavras abertas aprendidas.

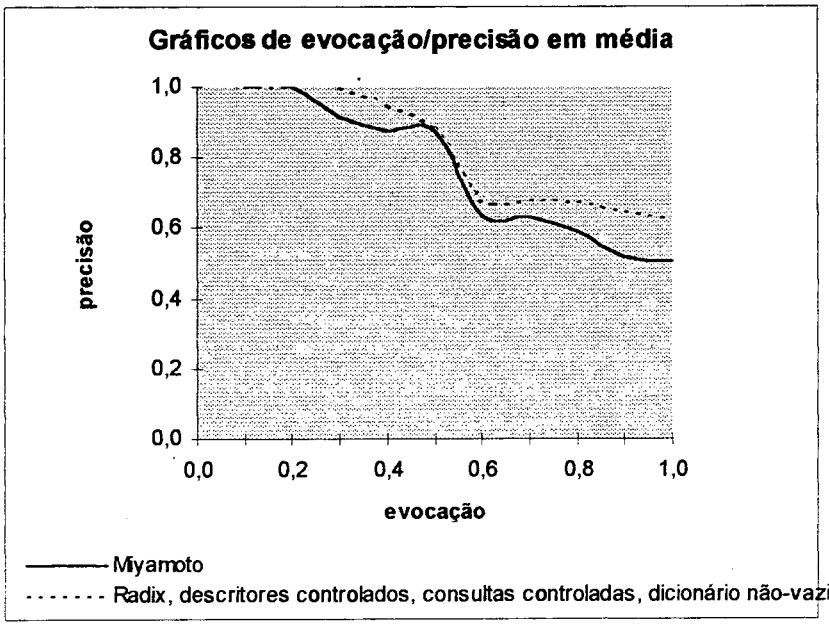


Figura 5.8: Gráficos de evocação/precisão em média para o modelo de Miyamoto e para o sistema Radix considerando descritores controlados, consultas controladas e iniciando com um conjunto de palavras abertas aprendidas.

Aplicou-se o teste de Wilcoxon para determinar se as diferenças da precisão nos níveis da evocação entre o sistema de Miyamoto e o sistema Radix são estatisticamente significativas ou não. Este teste compara duas variáveis e considera as seguintes hipóteses:

Hipótese nula: as variáveis são iguais;

Hipótese alternativa: as variáveis são diferentes.

A figura 5.9 mostra o resultado do teste, tanto nos níveis da evocação, quanto considerando o total de observações.

Wilcoxon Matched Pairs Test

evocação	Miyamoto vs Radix, descritores controlados, consultas controladas, dicionário não-vazio			
	N	T	Z	p-level
0,1	8	0.0	--	--
0,2	8	0.0	--	--
0,3	8	0.0	--	--
0,4	8	1.0	1.0690	.2851
0,5	8	3.0	0.0000	1.0000
0,6	8	6.0	.4045	.6858
0,7	8	6.0	.4045	.6858
0,8	8	7.0	.7338	.4631
0,9	8	5.0	1.1531	.2489
1.0	8	3.0	1.5724	.1159
Total	80	165.0	2.6394	.008310

Figura 5.9: Teste de Wilcoxon comparando o modelo de Miyamoto e o sistema Radix considerando descritores controlados, consultas controladas e iniciando com um conjunto de palavras abertas aprendidas.

A figura mostra que se pode rejeitar a hipótese nula, por exemplo, a um nível de significância de 5%, considerando o total das observações. Portanto, considerando também a figura 5.8, pode-se afirmar que existem evidências de que o sistema Radix, utilizando um vocabulário controlado, apresenta uma maior precisão nos níveis de evocação do que o sistema de Miyamoto.

5.2.3.2 Radix com descritores controlados vs Radix com descritores não-controlados

As figuras 5.10 e 5.11 mostram a evocação e o fallout do sistema Radix para consultas que não seguem um vocabulário controlado considerando os seguintes casos:

- descritores controlados e dicionário vazio: os descritores seguem o vocabulário controlado (anexo 3) e o analisador morfológico inicia com um dicionário vazio;
- descritores controlados e dicionário não-vazio: os descritores seguem o vocabulário controlado (anexo 3) e o analisador morfológico inicia com um dicionário de 158 palavras da classe aberta aprendidas (anexo 4);

- descritores não-controlados e dicionário vazio: os descritores não seguem um vocabulário controlado e o analisador morfológico inicia com um dicionário vazio;
- descritores não-controlados e dicionário não-vazio: os descritores não-seguem o vocabulário controlado e o analisador morfológico inicia com um dicionário de 158 palavras da classe aberta aprendidas (anexo 4).

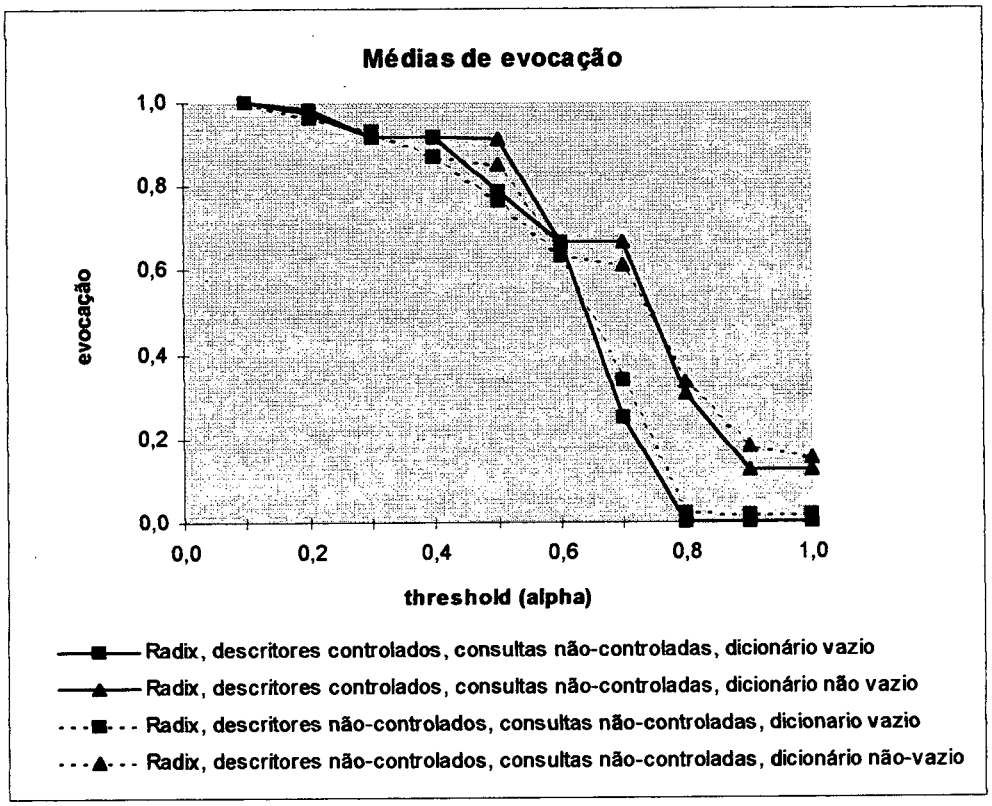


Figura 5.10: Médias de evocação nos níveis do threshold para o sistema Radix considerando descritores controlados e descritores não-controlados, consultas não-controladas e iniciando sem palavras abertas aprendidas e com um conjunto de palavras abertas aprendidas.

A figura 5.10 mostra a importância do dicionário para a evocação. Para valores de α entre 0.6 e 1.0, os dois casos com dicionário não-vazio apresentam uma evocação bem maior do que os casos que iniciam com um dicionário vazio. Por outro lado não se pode observar muita diferença nas médias de evocação considerando, por um lado, os dois casos com descritores controlados e, por outro lado, os dois casos com descritores não-controlados.

Para as médias dos fallout nos níveis do threshold α não se pode observar muitas diferenças para os quatro casos em consideração. Para valores de α entre 0.1 e 0.4 pode-se notar médias de fallout um pouco maiores para os casos com descritores não-controlados e para valores de α entre 0.6 e 0.8 pode-se observar médias de fallout um pouco maiores para os casos com dicionário não-vazio.

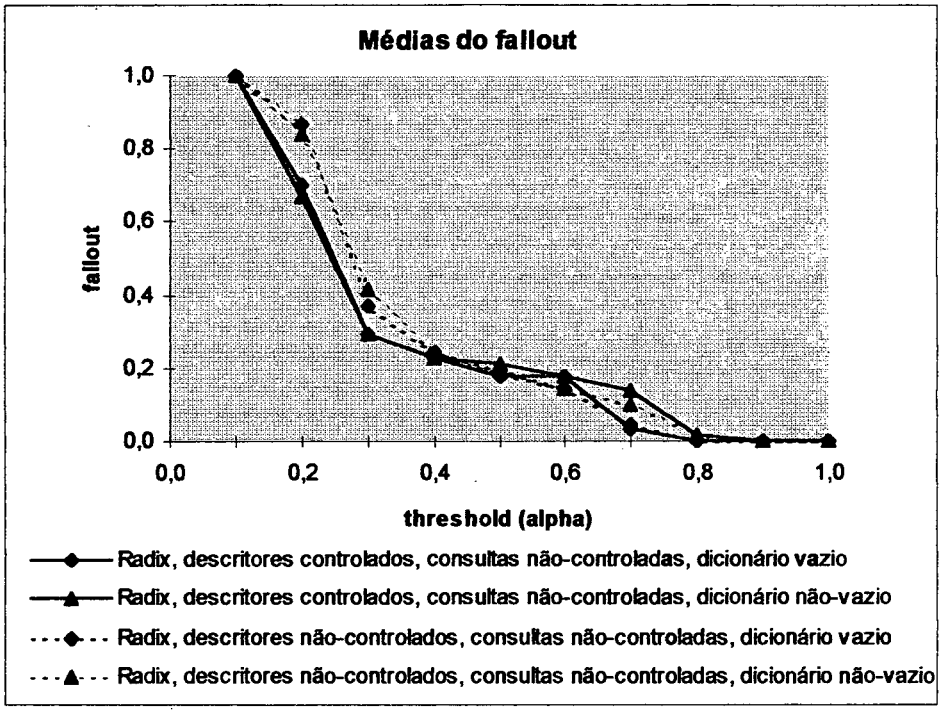


Figura 5.11: Médias de fallout nos níveis do threshold para o sistema Radix considerando descritores controlados e descritores não-controlados, consultas não-controladas e iniciando sem palavras abertas aprendidas e com um conjunto de palavras abertas aprendidas.

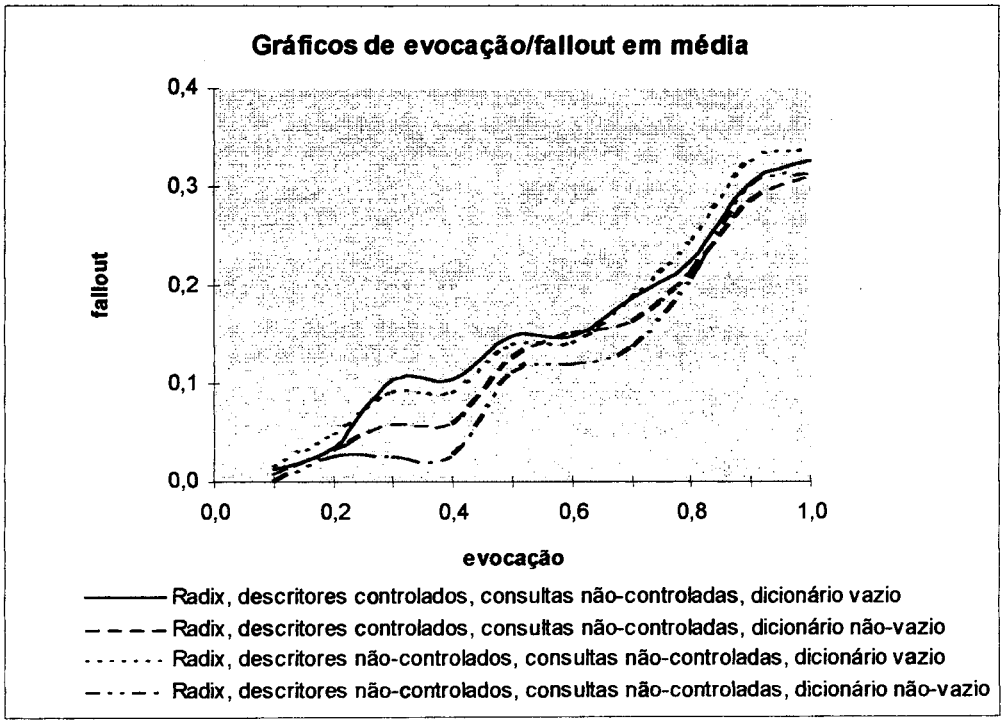


Figura 5.12: Gráficos de evocação/fallout em média para o sistema Radix considerando descritores controlados e descritores não-controlados, consultas não-controladas e iniciando sem palavras abertas aprendidas e com um conjunto de palavras abertas aprendidas.

A figura 5.12 mostra os gráficos de evocação/fallout para os quatro casos em consideração. Os dois casos com dicionário não-vazio apresentam valores de fallout nos níveis de evocação menores do que os dois casos com dicionário vazio. Além disso pode-se observar um melhor desempenho do sistema Radix com descritores não-controlados e dicionário não-vazio para valores de evocação entre 0.1 e 0.8 e um melhor desempenho do sistema Radix com descritores controlados e dicionário não-vazio para valores de evocação entre 0.8 e 1.0.

5.2.3.3 Miyamoto vs Radix com descritores não-controlados e consultas não-controladas

A figura 5.13 mostra as médias de evocação e precisão nos níveis do threshold para o sistema Miyamoto e o sistema Radix com descritores não-controlados, consultas não controlados e iniciando com o dicionário de 158 palavras aprendidas (anexo 4).

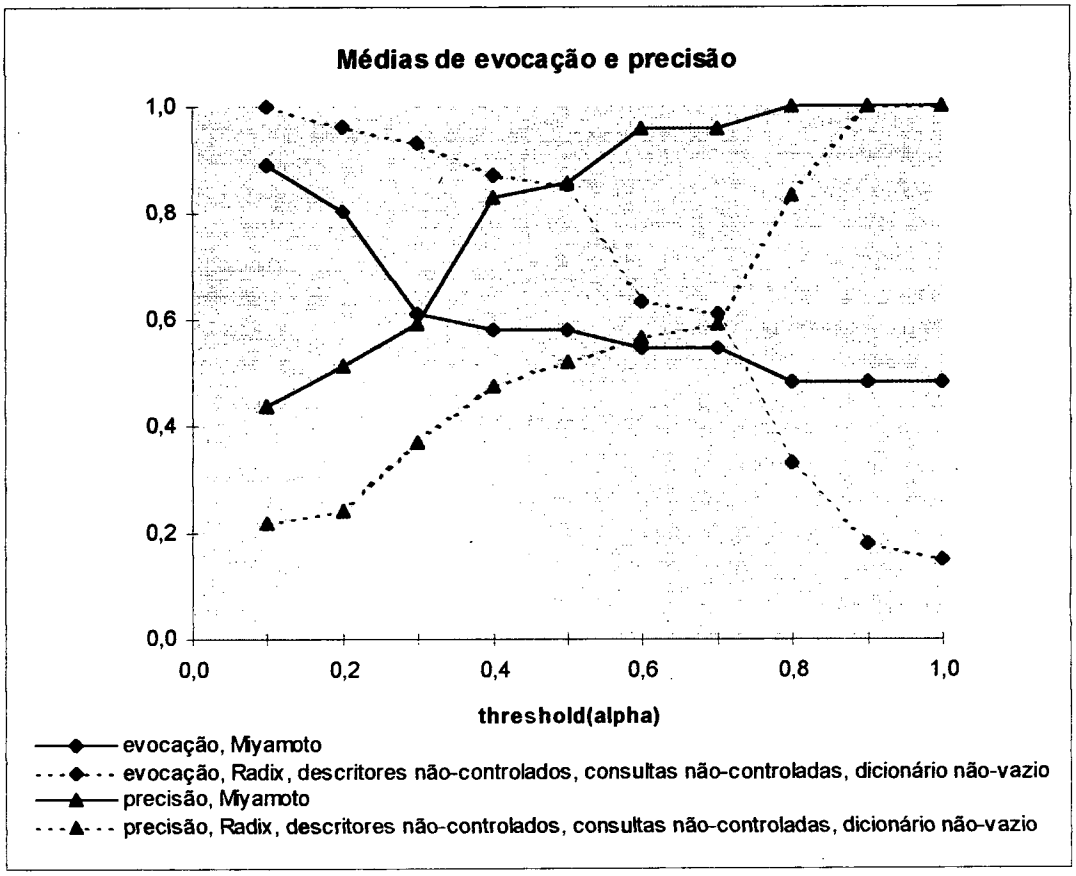


Figura 5.13: Médias de evocação e precisão nos níveis do threshold para o modelo de Miyamoto e para o sistema Radix considerando descritores não-controlados, consultas não-controladas e iniciando com um conjunto de palavras abertas aprendidas.

A figura 5.13 mostra uma maior precisão nos níveis do threshold para o sistema Miyamoto do que para o sistema Radix. A evocação é maior para o sistema Miyamoto

nos valores do threshold entre 0.8 e 1.0, enquanto o sistema Radix apresenta maior evocação nos valores do threshold entre 0.1 e 0.7.

A figura 5.14 mostra os gráficos de evocação/precisão para o sistema Miyamoto e o sistema Radix com descritores não-controlados, consultas não-controladas e iniciando com um dicionário de palavras da classe aberta aprendidas (anexo 4). O sistema Miyamoto apresenta maiores valores de precisão para valores de evocação entre 0.1 e 0.3 e entre 0.4 e 0.5, enquanto o sistema Radix apresenta maiores valores de precisão para valores de evocação entre 0.3 e 0.4 e entre 0.5 e 1.0.

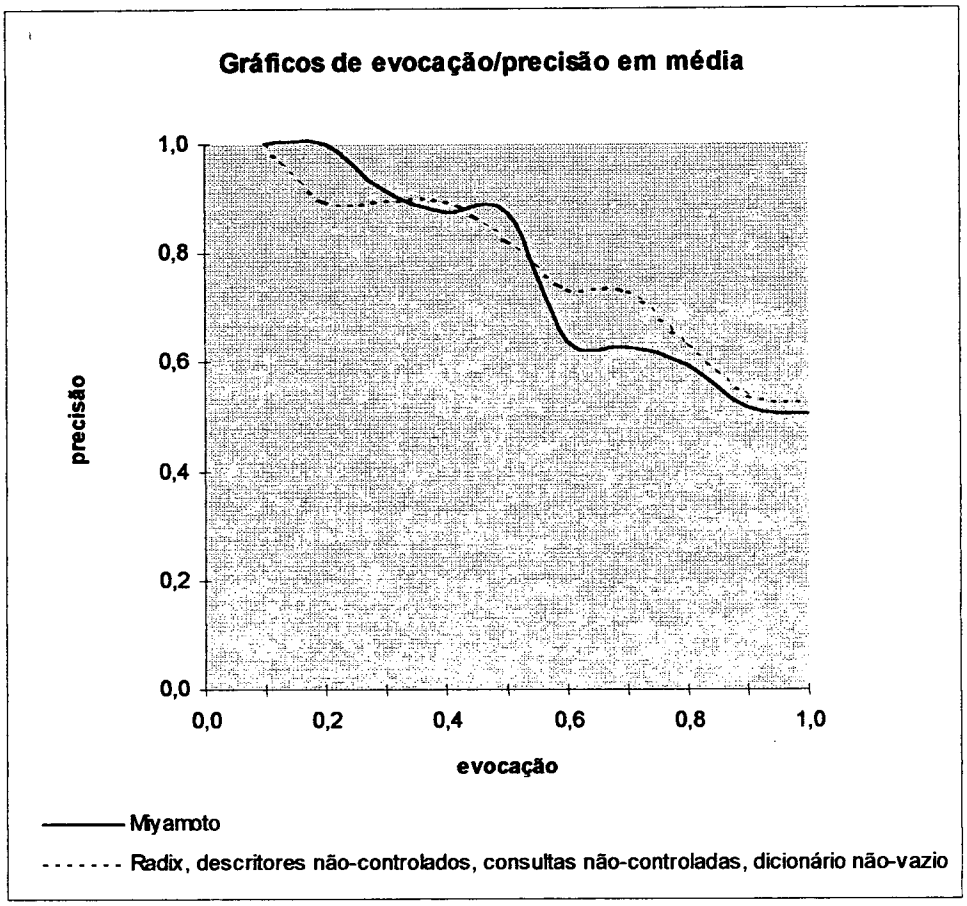


Figura 5.14: Gráficos de evocação/precisão em média para o modelo de Miyamoto e para o sistema Radix considerando descritores não-controlados, consultas não-controladas e iniciando com um conjunto de palavras abertas aprendidas.

Para verificar se existem diferenças estatisticamente significativas entre os valores de precisão nos níveis da evocação do sistema de Miyamoto e do sistema Radix com descritores não controlados, consultas não controladas e iniciando com o dicionário de 158 palavras aprendidas, aplicou-se o teste de Wilcoxon tanto para os valores de precisão obtidos nos níveis de evocação quanto ao total de observações. Assim testou-se a hipótese nula que as variáveis sejam iguais contra a alternativa que as variáveis sejam diferentes.

Os resultados obtidos e apresentados na figura 5.15 mostram que não se pode rejeitar a hipótese nula, nem em um nível de evocação, nem considerando o total de

observações, por exemplo, para um nível de significância de 5%. Portanto, tem-se que aceitar a hipótese nula e pode-se afirmar que não existem evidências para diferenças de valores de precisão nos níveis de evocação entre o sistema Miyamoto e o sistema Radix nas condições consideradas.

Cabe lembrar que a determinação da precisão nos níveis do threshold para o sistema Radix com um vocabulário não controlado considera somente as consultas que tiveram um retorno. Isto leva a altos valores de precisão e baixos valores de evocação para altos valores de α, o que não reproduz o desempenho real.

Wilcoxon Matched Pairs Test

	Miyamoto vs Radix, descritores não-controlados, consultas não-controladas, dicionário não-vazio			
evocação	N	T	Z	p-level
0,1	8	0.0	--	--
0,2	8	0.0	1.60357	.10882
0,3	8	3.0	.73030	.46521
0,4	8	6.0	.40452	.68583
0,5	8	4.0	.94388	.34524
0,6	8	5.0	1.15311	.24887
0,7	8	5.0	1.15311	.24887
0,8	8	10.0	.67612	.49897
0,9	8	11.0	.50709	.61209
1,0	8	10.0	.67612	.49897
Total	80	557.0	.77709	.43711

Figura 5.15: Teste de Wilcoxon comparando o modelo de Miyamoto e o sistema Radix considerando descritores não-controlados, consultas não-controladas e iniciando com um conjunto de palavras abertas aprendidas.

5.3 Considerações finais do capítulo

Para a avaliação do modelo proposto fez-se uma implementação em Smalltalk, chamada Radix, e comparou-se o modelo proposto em testes com o modelo de Miyamoto. Para esta comparação testou-se o sistema Radix em seis diferentes casos e fez-se três diferentes comparações:

1. Miyamoto vs Radix com vocabulário controlado: o sistema Miyamoto apresentou maior precisão nos níveis do threshold, enquanto Radix apresentou maior evocação. Radix mostrou maior precisão nos níveis de evocação, estatisticamente significativa, ao nível de significância de 5%. Isto pode ser interpretado como um melhor desempenho para o thesaurus difuso determinado através do índice difuso que para o thesaurus difuso considerando simplesmente co-ocorrências de descritores;

2. Radix com consultas que não seguem o vocabulário controlado: nestas condições mostrou-se a grande importância do dicionário de palavras abertas, enquanto o tipo de descritores não influenciou muito nos resultados obtidos;

3. Miyamoto vs Radix sem vocabulário controlado: nestas condições o sistema Miyamoto apresentou um melhor desempenho, mas uma maior precisão nos níveis da evocação não pôde ser verificada.

Além disso deve ser salientado que a implementação do sistema Radix é mais complexa. Em termos de complexidade de algoritmos, o modelo proposto apresenta, teoricamente, uma maior complexidade. Observando que a determinação de similaridade entre expressões é, em complexidade, equivalente ao problema de determinação do caminho de menor custo em um grafo, verifica-se uma complexidade não-polinomial para o algoritmo implementado. Mas para *n* documentos, considerando condicões como - existe um número máximo de palavras abertas em expressões, existe um número máximo de descritores para os documentos, existe um dicionário fixo de palavras abertas, um Corpus, considerado pelo analisador morfológico - determina-se para o sistema Radix uma complexidade de $O(n^2)$, igual a complexidade do modelo de Miyamoto.

6. CONCLUSÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

A perspectiva da interconexão entre computadores, através de redes com cobertura internacional, trouxe um significado ainda maior para a área de recuperação automática de documentos. Observa-se, por exemplo, a importância de sistemas como 'Alta Vista', 'Yahoo' ou 'Webcrawler' para pesquisas bibliográficas na Internet.

A utilização da teoria de conjuntos difusos para a recuperação de informações tem como vantagem o fato de que as relações difusas são mais expressivas do que as relações não-difusas, e a construção das relações difusas é mais realística. A resposta do sistema difuso ainda pode diferenciar entre os documentos recuperados via um grau de pertinência. Este grau de pertinência pode servir para o usuário como uma diretriz para a relevância dos documentos recuperados.

A qualidade de um sistema de recuperação de documentos depende da qualidade do thesaurus utilizado. Utilizando-se um vocabulário controlado, o modelo proposto no capítulo 4 mostrou nos testes uma maior evocação e uma menor precisão nos níveis do threshold do que o modelo de Miyamoto, bem como uma maior precisão nos níveis da evocação. Com isso o método de construção de um thesaurus difuso para a língua portuguesa através de um índice difuso, apresentado no item 4.3.1, constitui uma boa alternativa à construção do thesaurus difuso somente considerando co-ocorrências.

Utilizando-se um vocabulário não-controlado, os testes indicam um desempenho melhor nos níveis do threshold para o modelo de Miyamoto do que para o modelo proposto, no entanto, não foi possível constatar uma diferença estatisticamente significativa da precisão nos níveis da evocação entre os dois modelos. Sendo assim, o modelo proposto pode ser visto como uma alternativa em casos em que é difícil estabelecer um vocabulário controlado e consistente. Considerando, por exemplo, a área de programação orientada a objetos, é difícil, se não impossível, preestabelecer um vocabulário que sirva para a recuperação de classes de uma base de classes, uma vez que tudo pode ser objeto. Por outro lado, não se pode esperar a construção de um vocabulário consistente na inclusão de novas classes por parte do usuário. Os resultados apresentados no capítulo 5 mostram que, neste caso, a utilização de um vocabulário livre tanto nas consultas quanto nos descritores pode ser uma possível solução.

Considerando os resultados dos testes deve-se ainda salientar que o número de consultas e a base de documentos utilizados foram pequenos. Por outro lado, existem

também várias possibilidades de melhorar o sistema proposto. Por exemplo poder-se-ia utilizar outras adaptações do 'e' lógico difuso no tratamento das certezas pelo analisador morfológico ou a atribuição de penalidades mais variadas na análise das expressões. Pode-se, ainda, permitir operadores lógicos difusos, 'e' e 'ou', nas consultas, aplicando estes operadores nos componentes da relação *R* antes da aplicação do filtro, semelhante ao procedimento de Shyi-Ming Chen e Jeng-Yih Wang [CHE 95].

Além disso, o procedimento de determinação de similaridades entre expressões, pode, também, servir para criar extensões de modelos difusos como, por exemplo, o modelo de Murai, Miyakoshi e Shimbo [MUR 88;89] ou o modelo de Nomoto, Wakayama, Kirimoto; Ohashi e Kondo [NOM 90].

Para trabalhos futuros poderia-se pensar em algoritmos que geram automaticamente as relações do thesaurus *T* e da relevância dos descritores para os documentos *V* a partir de textos inteiros ou resumos. O modelo apresentado considera tanto uma análise morfológica "soft" através da relação difusa entre palavras *PP*, quanto uma análise semântica "soft" através da análise de expressões e da construção do thesaurus difuso. Uma extração de informação de textos inteiros ou resumos necessita ainda uma análise sintática, na preferência também "soft". Esta análise sintática "soft" poderia ser feito, por exemplo, adaptando modelos clássicos de análise sintática ao modelo apresentado, considerando gramáticas difusas, modelos Hidden Markov possibilísticos ou redes neurais difusas. Este modelo de processamento "soft" de linguagem natural poderia também ser aplicada na busca direta de informações, não considerando thesaurus e relação de relevância de descritores para documentos.

7. REFERÊNCIAS BIBLIOGRÁFICAS

[AND 91] ANDERSON; S. R.. *A morphous morphology.* Department of Cognitive Science. The Johns Hopkins University, 1991.

[BAR 93] BARRON, J.J. Putting fuzzy logic into focus. *Byte.* abr., p.111-118, 1993.

[BEL 89] BELEW, R.K. Adaptive information retrieval. Proceedings of the twelfth Annual International ACM/Sigir Conference on Research and Development in Information Retrieval. Jun 25-28, *Special Issue of the Sigir Forum*, p.11-20, 1989.

[BRU 91] BRUZA, P.D.; VAN DER WEIDE, T.P. The modelling and retrieval of documents using index expressions. *ACM SIGIR Forum*, v.25, n.2, p.91-103, 1991.

[CAB 85] CABRAL, L.S. *Introdução à lingüística.* 6.ed. Porto Alegre: Globo, 1985.

[CHE 95] CHEN, SHYI-MING ; WANG, JENG-YIH. Document retrieval using knowledge-based fuzzy information retrieval techniques. *IEEE Transactions on Systems, Man and Cybernetics.* v. 25, n.5, p.793-803, 1995.

[COU 92] COULON, D.; KAYSER, D. *Informática e linguagem natural.* Brasília: CNPq, 1992.

[CRO 83] CROFT, W.B. Applications for information retrieval techniques in the office. Proceedings of the sixth Annual International ACM/Sigir Conference on Research and Development in Information Retrieval. *ACM SIGIR Forum.* v. 17, n.4, p.18-23, 1983.

[DIG 92] DIGITALK. *Smalltalk/V for Windows. Object oriented programming system. Tutorial and programming handbook.* Los Angeles: Digitalk Incorporated, 1992

[EAS 78] EASTMAN, C.; WEISS, S. A tree algorithm for nearest neighbor searching in document retrieval systems. *ACM SIGIR Forum.* v. 13, n.1, p.131-149, 1978.

[FAR 86] FARRENY, H.; PRADE, H. *Dealing with the vagueness of natural languages in man-machine communication.* In: KARWOWSKI, W.; MITAL, A. (Ed.) Applications of fuzzy set theory in human factors. Amsterdam: Elsevier Science, p. 71-133, 1986.

[FAV 95] FAVERI, C.B.de.. A inteligência artificial e a lingüística. I Encontro do Celsul, 13-14 Nov, Florianópolis, *Anais*, v.2, p.604-610, 1995.

[FRA 89] FRAKES, W.B.; GANDEL, P.B. Classification, storage and retrieval of reusable components. Proceedings of the twelfth Annual International ACM/Sigir Conference on Research and Development in Information Retrieval. Jun 25-28, *Special Issue of the Sigir Forum*, p.251-254, 1989.

[GRE 92] GREFENSTETTE, G. Use of syntactic context to produce term association list for text retrieval. Proceedings of the twelfth Annual International ACM SIGIR Conference on Reseach and Development in Information Retrieval. jun 25-28, p.89-97, 1989. *Special Issue of the SIGIR Forum.*

[HED 95] HEDBERG, S.R. The data gold rush. *Byte.* oct, p.83-88, 1995.

[HES 92] HESS, M. An incrementally extensible document retrieval system based on linguistic and logical principles. Fifteenth Annual International ACM/Sigir Conference on Research and Development in Information Retrieval. 21-24 Jun., Dinamarca, *Proceedings*, p.190-197, 1992.

[ILA 85] ILARI, R.; GERALDI, J.W. *Semântica*. 2.ed. São Paulo: Ática, 1985.

[JAR 85] JARKE, M.; TURNER, J.A.; STOHR, E.A.; VASSILOU, Y.; WHITE, N.H.; MICHIELSEN . A field evaluation of natural language for data retrieval. *IEEE Transactions on software engeneering*. v. se-11, n.1, p.97-113, 1985.

[KLI 95] KLIR, G.J; YUAN, B. *Fuzzy sets and fuzzy logic: theory and applications*. London: Prentice-Hall, 1995, 573 p.

[LAN 87] LANCASTER, F.W. *Construção e uso de tesauros: curso condensado*. Brasília: IBICT, 1987. 106p.

[LAR 93] LARSEN, H.L.; YAGER, R. The use of fuzzy relational thesauri for classificatory problem solving in information retrieval and expert systems. *IEEE Transactions on Systems, Man and Cybernetic*. v. 23, n.1, p.31-41, 1993.

[LOP 96] LOPES, G.P. *Combining natural language understanding and information retrieval for flexible hypertext navigation*. 13 th. Brazilian Symposium on Artificial Intelligence, SBIA'96. Curitiba, Brasil, out. 1996. In: BORGES, D.L.B. & KAESTNER, C.A.A. (Ed.) Advances in Artificial Intelligence (Lecture Notes in Artificial Intelligence; 1159). Berlin: Springer, 1996, p. 233. (e material fornecido pelo autor).

[LUK 95] LUKASHEVICH, N.V. Automated formation of an information-retrieval thesaurus on the contemporary sociopolitical life of Russia. *Automatic Documentation and Mathematical Linguistics*. v. 29, n.2, p.29-35, 1995.

[MIY 83] MIYAMOTO, S.; MIYAKE, T.; NAKAYAMA, K. Generation of a pseudothesaurus for information retrieval based on cooccurrences and fuzzy set operations. *IEEE Transaction on Systems, Man and Cybernetics*. v. smc-13, n.1, p.62-70, 1983.

[MIY 86] MIYAMOTO, S.; NAKAYAMA, K. Fuzzy information retrieval based on a fuzzy pseudothesaurus. *IEEE Transaction on Systems, Man and Cybernetics*. v. smc-16, n.2, p.278-282, 1986.

[MIY 89] MIYAMOTO, S. Two Approaches for information retrieval through fuzzy association. *IEEE Transactions on Systems, Man and Cybernetics*. v. 19, n.1, p.123-130, 1989.

[MIY 90] MIYAMOTO, S. Information retrieval based on fuzzy association. *Fuzzy Sets and Systems*. v. 38, p.191-205, 1990.

[MUR 88] MURAI, T.; MIYAKOSHI, M.; SHIMBO, M. A modeling of search oriented thesaurus use based on multivalued logical inference. *Information Sciences*. v. 45, p.185-212, 1988.

[MUR 89] MURAI, T.; MIYAKOSHI, M.; SHIMBO, M. A fuzzy document retrieval method based on two-valued indexing. *Fuzzy Sets and Systems*. v. 30, p.103-120, 1989.

[NAS 95] NAS EMPRESAS. O senhor on-line bate na burocracia. *Informática Exame*. dez., p.94-98, 1995.

[NOM 90] NOMOTO, K.; WAKAYAMA, S.; KIRIMOTO, T.; OHASHI, Y.; KONDO, M.. A document retrieval system based on citations using fuzzy graphs. *Fuzzy Sets and Systems.* v. 38, p.207-222, 1990.

[OUH 91] OUHALLA, J. *Functional categories and parametric variation.* London: Routledge, 1991.

[PEN 95] PENTEADO, S. O Brasil põe o pé na estrada. *Informática Exame.* nov., p.80-84, 1995.

[POL 87] POLLITT, S. Cansearch: An Expert System approach to document retrieval. *Information Processing & Management..* v. 23, n. 2, p.119-1138, 1987.

[RAD 81] RADECKI, T. Outline of a fuzzy logic approach to information retrieval. *Int. J.ournal Man Machine Studies.* v. 14, p.169-178, 1981.

[RAG 89] RAGHAVAN, V.V.; BOLLMANN, P.; JUNG, G.S. Retrieval system evaluation using recall and precision: problems and answers. Proceedings of the twelfth Annual International ACM SIGIR Conference on Reseach and Development in Information Retrieval. jun 25-28, p. 59-68, 1989. *Special Issue of the SIGIR Forum.*

[REI 94] REINHARDT, A. Building the data highway. *Byte.* março, p.46-74, 1994.

[ROB 82] ROBREDO, J. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico. *Ci. Inf. Brasília.* v. 11, n.1, p.3-18, 1982.

[RUS 95] RUSSEL, S.; NORVIG, P. *Artificial Intelligence: a modern approach*, New Jersey: Prentice Hall, 1995, 932p.

[SAL 75] SALTON, G. *Dynamic information and library processing*, New Jersey: Prentice Hall, 1975, 523p.

[SAL 83] SALTON; G.; McGILL, M.J. *Introduction to modern information retrieval.* Auckland: McGraw-Hill, 1983, 448p.

[SAL 89] SALTON, G.; SMITH, M. On the application of syntactic methodologies in automatic text analysis. Proceedings of the twelfth Annual International ACM SIGIR Conference on Reseach and Development in Information Retrieval. jun 25-28, p.137-150, 1989. *Special Issue of the SIGIR Forum.*

[SIL 89] SILVA, M.C.P.S. & KOCH, I.V. *Lingüística aplicada ao português: morfologia.* 5.ed. São Paulo: Cortez, 1989.

[SPA 96] SPARCK JONES, K.; GALLIERS, J.R. *Evaluating natural language processing systems: an analysis and review.* (Lecture Notes in computer science; 1083) Berlin: Springer, 1996, 225 p.

[STO 96] STORB, B.H; WAZLAWICK, R.S. A simple method for recognizing radicals and suffixes of unknown words. International Symposium on Multi-Technology Information Processing, ISMIP'96, 16-18 Dec, Hsinchu, Taiwan,*Proceedings*, p.589-593, 1996.

[VIL 94] VILLALVA, A. *Estruturas morfológicas. unidades e hierarquias nas palavras do português.* Dissertação de Doutoramento. Faculdades de Letras da Universidade de Lisboa. Lisboa, 1994. 423p.

[ZAD 65] ZADEH, L. A.. Fuzzy sets. *Information and Control*, v.8, n.3, p.338-353, 1965.

[ZAD 75] ZADEH, L. A.. The concept of a linguistic variable and its application to approximate reasoning-I. *Information Sciences*, v.8, p.199-249, 1975.

[ZEM 85] ZEMANKOVA-LEECH, M.; KANDEL, A. *Fuzzy relational data bases - a key to expert systems* Köln: TÜV Rheinland, 1985, 180 p.

[ZEN 85] ZENNER, R.B.R.C.; CALUWE, R.M.M.; KERRE, E.E. A new approach to information retrieval systems using fuzzy expressions. *Fuzzy Sets and Systems*, v.17, p.9-22, 1985.

8. ANEXOS

Anexo 1: A Method for Recognizing Radicals and Suffixes of Unknown Words

A Method for Recognizing Radicals and Suffixes of Unknown Words

Bernd Heinrich Storb - Raul Sidnei Wazlawick
{raul,bernd}@inf.ufsc.br

Universidade Federal de Santa Catarina, Centro Tecnológico
Departamento de Informática e Estatística
Curso de Pós-Graduação em Engenharia de Produção
UFSC-CTC-INE Cx.P. 476 - 88040-900 Florianópolis, SC Brasil

ABSTRACT

The authors suggest an approach to the learning of new words and new variations of known words in natural language morphological analysis. This method focuses on the determination of radicals and suffixes of words, and is implemented by splitting each word into a list of all possible radical/suffix pairs whereby the most promising of them is determined comparing certainty degrees. The calculus of the certainty degrees is based on combination of the certainty values obtained for each possible radical and suffix. Some combination techniques of these certainties are compared.

1. INTRODUCTION

Natural language processing systems based on stand-alone workstations are designed to enclose as much knowledge as possible, because it is difficult to acquire linguistic knowledge from common users. But a system used by a great number of users, like a network system, may get a better performance in learning new words, structures and linguistic significance.

Before discovering the meaning of a word it is necessary to find its morphology. This simplifies the process of word comprehension by providing recognition of many variations of one word (concept). This paper analyses some identification techniques and presents a comparative result.

It is possible to classify words in two categories: *open classes* and *closed classes*. Open classes are: verbs, nouns and adjectives. Closed classes cover the functional categories: articles, numerals, pronouns, adverbs, prepositions, conjunctions and interjections [4].

This classification of words has been largely used in natural language processing [2]. Almost all studies focus on open class word learning, since new words of a language almost always belong to these classes.

2. MORPHOLOGICAL ANALYZER

The morphological analyzer prototype's basic objects (implemented in Smalltalk [1]) are the *words*, its *radicals* and *suffixes*, and the *dictionaries*, which allow access to information about these objects. Each word has as attributes its own text and a value between 0 and 1 representing the system's certainty about it. Closed class words have certainty always set to 1. Open class words have also their radical and suffix as attributes. The dictionary of open class words associates to each word a table representing possible divisions of the word into radical/suffix pairs and the certainty degree assigned to each possible radical/suffix pair.

The learning process consists in a successive elimination of the pairs with a very low certainty value, until only one pair remains. Then the radical and suffix pair reach the status of *learnt* and the word certainty is set to 1.

If the lexical analyzer finds an unknown word, its first action is to create a table with all possible radical/suffix pairs, their certainties, and the combination of these values. For example, the word 'amigo' (friend) would be splitted in six radical/suffix pairs like the ones shown in the table 1.

Table1: Example of the process of splitting of the word "amigo" into all possible radical/suffix pairs; radical certainty (rc); suffix certainty (sc) and pair certainty.

Radical	Rad. Cert. *rc*	Suffix	Suff. Cert. *sc*	Pair certainty: min{*rc*,*sc*}
	1.0	amigo	0.0	0.0
a	0.2	migo	0.1	0.1
am	1.0	igo	0.3	0.3
ami	0.5	go	0.1	0.1
amig	0.8	o	1.0	0.8
amigo	0.3		1.0	0.3

Each line represents a possible division of the word. The system determines a degree of certainty (*rc* and *sc*) for each possible radical and suffix. In this example, the certainty of the combination is calculated as the minimum between both certainties. The most promising pair is "amig/o". Other combination functions will be discussed later.

The certainty of a learnt radical is 1. The certainty of a radical being learned is composed by a combination of all certainties of suffixes, that has been observed to occur with. For example, in Portuguese, the words "amigo", "amiga", "amigos" and "amigas" share the same radical. The same is true for suffixes. Therefore, the certainty of a radical is proportional to the number of its suffixes, and the certainty of a suffix is proportional to the number of its radicals.

3. CERTAINTY FUNCTIONS

Three different functions were used in order to combine the certainty of a radical and a suffix to form word certainty values:

a) *Product with reducers*: the certainty of a radical/suffix pair is calculated in the same way as the probability of an occurrence of two independent events. The result, that corresponds to the product of both certainties, is reduced if the size of the radical or suffix is very small:

$$certainty_{pair} = certainty_{radical} * certainty_{suffix} * \alpha_{|radical|} * \alpha_{|suffix|}$$

The reducers are defined by the function $\alpha_x = (12+x)/16$, if $0 \leq x \leq 3$, otherwise $\alpha_x = 1$. Moreover, $|radical|$ and $|suffix|$ denote the size of the radical and suffix.

b) *Weighted Average with reducers*: the certainty of the pair is a weighted average of the radical and suffix certainties. The value β, $0 \leq \beta \leq 1$, defines the weight of each certainty:

$$certainty_{pair} = (\beta * certainty_{radical} * \alpha_{|radical|}) + ((1-\beta) * certainty_{suffix} * \alpha_{|suffix|}).$$

c) *Minimum with reducers*: the certainty of the pair is the minimum between the certainty of the radical and suffix: $certainty_{pair} = \min \{certainty_{radical} * \alpha_{|radical|}, certainty_{suffix} * \alpha_{|suffix|}\}$ of

4. DELETION OF RADICAL/SUFFIX PAIRS WITH LOW CERTAINTY

Radical/suffix pairs with low certainty value will be removed from the table until only one pair remains and the word is considered learnt. In this process, two approaches were considered: *percentage* and *average/deviation*. These approaches depend on a given value, called *evalCoeficient*, which may vary from 0 to 1.

In the first approach (percentage), every pair with certainty below a percentage of the highest certainty is removed from the table, that is, the system removes the pairs that satisfy: $certainty_{pair} < \max\{ certainty_{pair(i)} \mid pair(i) \in table \} * evalCoeficient$. The second approach (average/deviation) removes the pairs that satisfy: $certainty_{pair} < average_{table} - (evalCoeficient * standardDeviation_{table})$.

5. TESTS AND RESULTS

Five same-subject texts were selected to test those rules defined above. The texts are: *A* (156 words), *B* (221 words), *C* (305 words), *D* (647 words), and *E* (1190 words). In the first test, the texts were analyzed with empty dictionaries, that is, no word, radical or suffix were previously known. Each text was analyzed varying three parameters: (a) certainty functions, (b) evaluation approaches and (c) different values for *evalCoeficient*. Following, each text was analyzed using four dictionaries: dictionary 1 (77 learnt words), dictionary 2 (128 learnt words), dictionary 3 (168 learnt words) and dictionary 4 (484 learnt words).

Table 2: Results obtained for the text E (1190 words), varying the parameters (certainty function, deletion approach, evaluation coefficient) and using dictionary 3 (168 words, not necessarily containing the words of text 3).

certainty function	deletion approach	evaluation coefficient	% right answers	% right answers between the three best answers
product	percentage	0,25	58,42	83,14 (+)
		0,5	57,30	75,28
		0,75	56,17	62,92
	average/dev.	0,5	59,55	67,41
		0,75	59,55	70,78
		1,0	66,29 (*)	82,02 (+)
weighted average (β = 0,5)	percentage	0,25	34,83	80,90 (+)
		0,5	34,83	80,90 (+)
		0,75	35,96	66,29
	average/dev.	0,5	41,57	55,06
		0,75	43,82	60,67
		1,0	47,19	69,66
minimum	percentage	0,25	56,17	87,64 (+)
		0,5	56,17	87,64 (+)
		0,75	55,05	68,53
	average/dev.	0,5	59,55	67,41
		0,75	53,93	67,41
		1,0	65,16 (**)	82,14 (+)

The best results to text E with dictionary 3 (table 2) were obtained using product function and average/deviation with evaluation coefficient 1,0 (*). The second best result was the method using minimum function and average/deviation with coefficient 1,0 (**). Both methods gave more than 65% right answers, and more than 80% if the three best answers were considered (+). It can be concluded that the technique of average/deviation with coefficient 1,0 using minimum or product is the best. The weighted average seems to be the worst technique.

The following graphic (figure 1) shows the results obtained in the average for texts A, B, C, D and E being analyzed using method (**), comparing the results obtained with the empty dictionary and dictionaries 1, 2, 3 and 4:

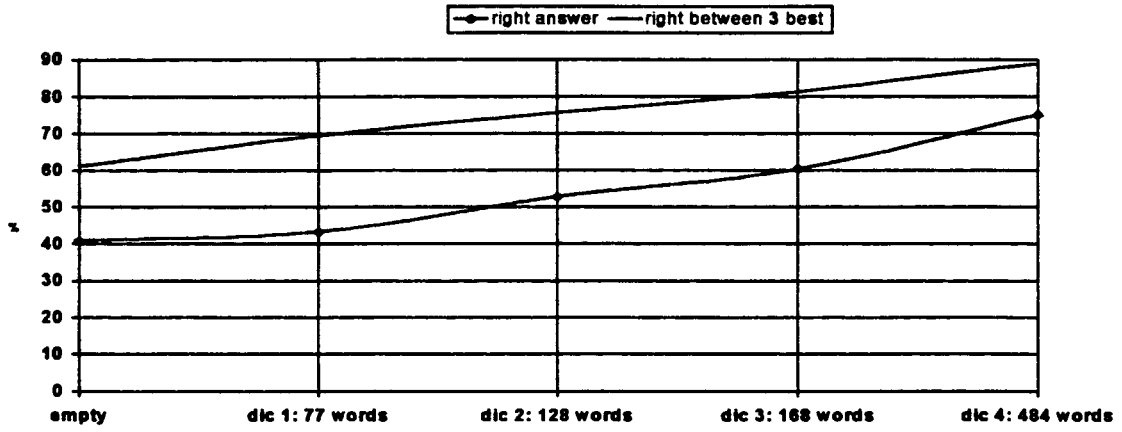


Figure 1: Relation between the dictionaries and the averages of results obtained for the different texts.

The increasing number of right answers let us presume that it is possible to determine radical and suffix of unknown words using certainty combinations of the possible radicals and suffixes. Of course, these results are not good enough if compared to methods using a large built-in lexicon, but the developed method do not use any previous knowledge about the language, and can be suited to many of them.

6. CONCLUSIONS AND FURTHER WORK

The prototype developed for morphological analysis seems to be valid for the recognition of radicals and suffixes of Portuguese words. The tests gave good results even with quite small dictionaries. The increasing observed in figure 1 shows that enhancing the dictionary provides better results.

By the results, we suppose that this model might be combined with other methods of morphological analysis, such as the method of McFETRIDGE and VILLAVICENCIO for verb determination [3], or with functionalist part-of-speech-tagging without previous information about lexical categories [5]. This is left to a future work.

The complexity of the algorithm used in the analysis is not hard. Time complexity can be determined considering the number of accesses to the dictionaries. Space complexity may be measured as the size of the dictionaries. Considering that Portuguese words are limited to a maximum size (say, the size of "inconstitucionalissimamente"), the time and space complexity are upper bounded to $O(n^2)$, where n is the number of different words ever read by the system. But the number of suffixes in Portuguese may be considered almost constant (even if it is very great), so that the arising of new

suffixes does not need to be considered if the system have enough training. If the number of suffixes is c, then the complexity of time and space may be measured by $O(c.n)$, that is $O(n)$. So, it can be considered that this system is viable as part of a morphological analyzer.

7. REFERENCES

[1] *Smalltalk/V 286. Tutorial and Programming Handbook*. Los Angeles: IBM, 1988.

[2] GROSZ, B.; APPELT, D.; MARTIN, P.; PEREIRA, F. TEAM: An Experiment in the Design of Transportable Natural Languages Interfaces. *Artificial Intelligence* **32**:173-243, 1987.

[3] McFETRIDGE, P. & VILLAVICENCIO, A. A Hierarchical Description of the Portuguese Verb. 12. *Brazilian Symposium on Artificial Intelligence - SBIA, 95*. 11-13 Set. Campinas, Brazil. Proceedings, p.302, 1995.

[4] OUHALLA, J. *Functional categories and parametric variation*. London: Routledge, 1991.

[5] PACHECO, H.; DILLINGER, M.; CARVALHO, M. de. Uma nova abordagem para a análise sintática do Português. *II Encontro para o Processamento Computacional de Português Escrito e Falado. (XIII Brazilian Symposium on Artificial Intelligence - SBIA, 96)* Curitiba, Brazil. Proceedings, p.51-60, 1996.

ANEXO 2: Lista dos documentos considerados

1. **Autores**: Shenoi, S.; Melton, A.
   Título: Proximity relations in the fuzzy relational database model.
   Palavras-chave: armazenamento de informação, recuperação de informação, banco de dados relacional difuso, relação de proximidade, relação de similaridade, composição max-min.
   Expressões: armazenamento de informação, recuperação de informação, banco de dados relacional difuso, relações de proximidade, relações de similaridade, composição max-min.

2. **Autores**: Akdag, H.; Bouchon, B.
   Título: Using fuzzy set theory in the analysis of structures of information.
   Palavras-chave: questionário, hierarquia, informação, incerteza, operador difuso, tomada de decisão.
   Expressões: questionários, hierarquia, informações, incertezas, operadores difusos, tomada de decisões.

3. **Autores**: Teixeira, A.R.; Spieguel, C.P.
   Título: Banco de dados do programa flora do CNPq, sobre plantas medicinais e farmacologia de produtos naturais.
   Palavras-chave: banco de dados, plantas medicinais, farmacologia de produtos naturais.
   Expressões: banco de dados, plantas medicinais, farmacologia de produtos naturais.

4. **Autores**: George, R.; Buckles, B.P.; Petry, F.E.
   Título: Modelling class hierarchies in the fuzzy object-oriented data model.
   Palavras-chave: hierarquia, relação de similaridade, banco de dados orientado a objetos.
   Expressões: hierarquia de classes, relações de similaridade, bases de dados orientadas a objetos.

5. **Autores**: Korth, H.F.; Silberschatz
   Título: Sistemas de banco de dados.
   Palavras-chave: banco de dados relacionai, dependência funcional, linguagem de consulta.
   Expressões: banco de dados relacional, dependência funcional, linguagem de consulta.

6. **Autores**: Bruza, P.D.; Van der Weide, T.P.
   Título: The modelling and retrieval of documents using index expressions.
   Palavras-chave: recuperação de documentos, expressão indexada, lógica.
   Expressões: recuperação de documentos, expressão indexada, lógica formal.

7. **Autores**: Buneman, P.; Ohori, A.
   Título: Polymorphism and type inference in database programming.
   Palavras-chave: herança, banco de dados, banco de dados orientado a objetos, polimorfismo.
   Expressões: herança, base de dados, base de dados orientada a objetos, polimorfismo.

8. **Autores**: Brown, K.S.
Título: Engenharia ecológica: novas perspectivas de seleção e manejo de plantas medicinais.
Palavras-chave: plantas medicinais, engenharia ecológica, fitofármacos, variação de plantas, seleção natural.
Expressões: plantas medicinais, engenharia ecológica, fitofármacos, variação de plantas, seleção natural.

9. **Autores**: Pires, J.M.
Título: Plantas ictiotóxicas: aspecto da botânica sistemática.
Palavras-chave: plantas ictiotóxicas, botânica sistemática, classificação taxonômica, fitoquímica.
Expressões: plantas ictiotóxicas, botáncia sistemática, classificação taxonômica, fitoquímica.

10. **Autores**: Diamond, P.; Kloeden, P.
Título: Caracterization of compact subsets of fuzzy sets.
Palavras-chave: conjunto compacto, conjunto difuso, conjunto normalizado.
Expressões: conjuntos compactos, conjuntos difusos normalizados.

11. **Autores**: Mattos, N.M.
Título: Abstraction concepts: basis for data and knowledge.
Palavras-chave: banco de dados orientado a objetos, hierarquia, herança, polimorfismo, especificação, generalização.
Expressões: banco de dados orientado a objetos, hierarquia, herança, polimorfismo, especificação, generalização.

12. **Autores**: Gottsberger, I.S.
Título: O cerrado como potencial de plantas medicinais e tóxicas.
Palavras-chave: cerrado, plantas medinais, plantas tóxicas, estudos fitosociológicos, levantamento de espécies vegetais.
Expressões: cerrado, plantas medinais, plantas tóxicas, estudos fitosociológicos, levantamento de espécies vegetais.

13. **Autores**: Ponsard, C.
Título: Fuzzy mathematical models in economics.
Palavras-chave: conjunto difuso, cálculo econômico, equilíbrio econômico.
Expressões: conjuntos difusos, cálculo econômico, equilíbrio econômico I.

14. **Autores**: Buckley, J.I.; Tucker, D.M..
Título: Second generation fuzzy expert system.
Palavras-chave: sistema especialista, inteligência artificial, tomada de decisão.
Expressões: sistemas especialistas, inteligência artificial, tomada de decisão.

15. **Autores**: Willaeys, D.; Moreau, A.; Asse, A.
Título: Processing subjective information for diagnostic assistance.
Palavras-chave: diagnóstico, sistema homem-máquina, relação difusa.
Expressões: diagnósticos, sistema homem-máquina, relações difusas.

16. **Autores**: Maia, L.F.J.; Costa Jr., I.
Título: Fuzzy information retrieval for decision support.
Palavras-chave: recuperação de informação, conjunto difuso, tomada de decisão, banco de dados relacional difuso.

Expressões: recuperação da informação, conjuntos difusos, tomada de decisões, banco de dados relacional difuso.

17. Autores: Inui, M.; Shoaff, W.; Fausett, L.; Schneider, M.
Título: The recognition of imperfect strings generated by fuzzy context sensitive grammars.
Palavras-chave: gramática sensível ao contexto, conjunto difuso, modelo de reconhecimento.
Expressões: gramáticas sensíveis ao contexto, conjuntos difusos, modelos de reconhecimento.

18. Autores: Cubero, J.C.; Medina, J.M.; Pons, O.; Vila, M.A.
Título: Fuzzy loss less decompositions in databases.
Palavras-chave: dependência funcional, dependência difusa, regra difusa, projeção difusa.
Expressões: dependências funcionais difusas, regras difusas, projeção difusa.

19. Autores: Yager, R.R..
Título: Some properties of fuzzy relationships.
Palavras-chave: conjunto difuso, relação difusa, associação difusa.
Expressões: conjunto difuso, relação difusa, associação difusa.

20. Autores: Dubois, D.; Prade, H.
Título: Measuring properties of fuzzy sets: a general technique and its use in fuzzy query.
Palavras-chave: conjunto difuso, consulta vaga, informação, banco de dados.
Expressões: conjuntos difusos, consulta vaga, informação, base de dados.

21. Autores: Torres, C.A.G.; Barreiros, E.L.; Kaplan, M.A.C.; Gottlieb, O.R.
Título: Alcalóides esteroidais como marcadores evolutivos da familia buxaceae.
Palavras-chave: marcadores evolutivos, quimiossistemática, evolução natural.
Expressões: marcadores evolutivos, quimiossistemática, evolução natural.

22. Autores: Tripathy, R.C.; Saxena, P.C.
Título: Multivalued dependencies in fuzzy relational databases.
Palavras-chave: banco de dados relacional difuso, dependência funcional.
Expressões: banco de dados relacional difuso, dependência funcional.

23. Autores: Murai, T.; Miyakoshi, M.; Shimbo, M.
Título: A fuzzy document retrieval method based on two-valued indexing.
Palavras-chave: recuperação de documentos, conjunto difuso, lógica.
Expressões: recuperação de documentos, conjuntos difusos, lógica modal.

24. Autores: Ogawa, Y.; Morita, T.; Kobayashi, K.
Título: A fuzzy document retrieval system using the keyword connection matrix and a learning method.
Palavras-chave: recuperação de informação, relação difusa, indexação, aprendizagem, avaliação do desempenho.
Expressões: recuperação de informações, relações difusas, indexação, aprendizagem, avaliação dos desempenhos.

25. Autores: Bosc, P.; Galibourg, M.; Hamon, G.
Título: Fuzzy querying with SQL: extensions and implementation aspects.
Palavras-chave: banco de dados relacional, linguagem de consulta, condição difusa, indexação.

Expressões: banco de dados relacional, linguagem de consulta, condição difusa, indexação.

26. Autores: Rosenkranz, A; Jurkiewicz, A; Corrado, A.
Título: Situação dos biotérios brasileiros: fator limitante de estudos farmacodinâmicos e toxicológicos de produtos naturais.
Palavras-chave: biotério, estudos farmacodinâmicos, estudos toxicológicos, produtos naturais.
Expressões: biotério, estudos farmacodinâmicos, estudos toxicológicos, produtos naturais.

27. Autores: Smith. J.M.; Smith, D.C.P.
Título Database abstractions: aggregation and generalization.
Palavras-chave: banco de dados relacional, agregação, generalização.
Expressões: banco de dados relacional, agregação, generalização.

28. Autores: Buckles, B.P.; Petry, F.E.; Pillai, J.
Título: Network data models for representation of uncertainty.
Palavras-chave: banco de dados difuso, linguagem de consulta, conjunto difuso.
Expressões: base de dados difusa, linguagem de consulta, conjunto difuso.

29. Autores: Saade, J.J.; Schwarzlander, H.
Título: Ordering fuzzy sets over the real line: an approach based on decision making under uncertainty.
Palavras-chave: conjunto difuso, critério de decisão, tomada de decisão, incerteza.
Expressões: conjuntos difusos, critérios de decisão, tomada de decisão, incertezas.

30. Autores: Shenoi, S.; Melton, A.; Fan, L.T.
Título: An equivalence classes model of fuzzy relational databases.
Palavras-chave: banco de dados relacional difuso, dependência funcional.
Expressões: banco de dados relacional difuso, dependência funcional.

31. Autores: Ammar, S.
Título: Determining the 'best' decision in the presence of imprecise information.
Palavras-chave: conjunto difuso, tomada de decisão, classificação difusa.
Expressões: conjunto difuso, tomada de decisão, classificação difusa.

32. Autores: Gottlieb, O.R.;Stefanello, M.E.A.
Título: Avaliação estatística de plantas medicinais.
Palavras-chave: avaliação estatística, plantas medicinais.
Expressões: avaliação estatística, plantas medicinais.

33. Autores: Petrovick, R.R.; Mello, J.C.P.
Título: The use of fatorial design in the galenical development of phytotherapics.
Palavras-chave: análise fatorial, fitoterápicos, desenvolvimento galênico.

34. Autores: Mulani, J.; Bahulkar, A.
Título: A graphical navigator for viewing databases.
Palavras-chave: banco de dados orientado a objetos, linguagem de consulta, banco de dados visualizado.
Expressões: base de dados visualizada orientada a objetos, linguagem de consulta.

35. Autores: Miyamoto, S.
Título: Information retrieval based on fuzzy associations.
Palavras-chave: recuperação de informação, armazenamento de informação, associação difusa.
Expressões: armazenamento de informação, recuperação de informação, associação difusa.

36. Autores: Bardossy, A.; Bogardi, I.; Kelly, W.E.
Título: Geostatistics utilizing imprecise (fuzzy) information.
Palavras-chave: geoestatística, conjunto difuso, revestimento de solo.
Expressões: geoestatística, conjuntos difusos, revestimento de solo.

Anexo 3: Lista dos descritores controladas em ordem alfabética

agregação
análise fatorial
aprendizagem
armazenamento de informação
associação difusa
avaliação do desempenho
avaliação estatística
banco de dados
banco de dados difuso
banco de dados orientado a objetos
banco de dados relacional
banco de dados relacional difuso
banco de dados visualizado
biotério
botânica sistemática
cálculo econômico
cerrado
classificação difusa
classificação taxonômica
composição max-min
condição difusa
conjunto compacto
conjunto difuso
conjunto normalizado
consulta vaga
critério de decisão
dependência difusa
dependência funcional
desenvolvimento galênico
diagnóstico
engenharia ecológica
equilíbrio econômico
especificação
estudos farmacodinâmicos
estudos fitosociológicos
estudos toxicológicos
evolução natural
expressão indexada
farmacologia de produtos naturais
fitofármacos
fitoquímica
fitoterápicos
generalização
geoestatística
gramática sensível ao contexto
herança
hierarquia
incerteza
indexação
informação
inteligência artificial
levantamento de espécies vegetais
linguagem de consulta
lógica
marcadores evolutivos
modelo de reconhecimento
operador difuso
plantas ictiotóxicas
plantas medicinais
plantas tóxicas
polimorfismo
produtos naturais
projeção difusa
questionário
quimiossistemática
recuperação de documentos
recuperação de informação
regra difusa
relação de proximidade
relação de similaridade
relação difusa
revestimento de solo
seleção natural
sistema especialista
sistema homem-máquina
tomada de decisão
variação de plantas

ANEXO 4: DICIONÁRIO

Exemplos de palavras com sufixos nominais, de terminologia científica e adverbiais

aceitação
aceitável.
álgebra
algoritmo
alternante
alternativa
altivez
ambíguo
aprendiz
armazém
aromático
artificial
arvoredo
associada
astronomia
atualmente
avaliado
baseada
beleza
boiada
boiadas
brancura
calvície
campista
casarão
casarões
cemitério
classificados
compactação
comparado
compositor
confiante
conjunção
consolador
consulado
consultor
corrosivo
criterioso
decidido
dedutível
desenvolvido
desperdício
detenção
difusão
distributivo
documentação
doutorando
duradouro
economia
equilibrismo
equivalência
escolar
escuridão
esferóide
especial
esperança
expresso
extensivo
fechado
felizardo
ferimento
ferrugem
ferrugens
filtro
físico
físicos
formalismo
gastrite
gênero
gentil
gentís
geografia
geral
gigantesco
gorduroso
gritaria
hereditário
homenagem
imperial
imponência
incerto
índex
informativo
instantâneo
leiteria
leiterias
ligado.
limites
linguado
lisonjeiro
lisonjeiros
logismo
maldade
maquinista
marginalidade
mecânico
menina
meninas
menino
meninos
modelismo
movediço
mulherio
negrume
neurose
normalista.
objetivo
operação
ordenação
oriente
particularidade
patusco
paulada
poligonal
pontudo
preferido
princípio
processado
produtor
produtores
prognose
projeto
prosaico
provisório
próximo
questão
quietude
ramagem
recente
recuperado
redentor
registrado
regrado
relacionado
relutância
revestrés
rodoviário
romantismo
satisfatório
sedento
seletivo
sensibilidade
sindicato
sistemática
solúvel
substituto
suportável
teorema
tipografia
traição
trânsito
tristonho
útil
vagaroso
varicela
velhice
velhices
vítreo

ANEXO 5: Consultas para os testes

(C_i ~ consulta controlada; C'_i ~ consulta não-controlada)

1: Informações sobre banco de dados.

C_1 = 1.0 / banco de dados
C'_1= 1.0/ base de dados

2: Quase todas as informações sobre banco de dados relacional.

C_2 = 1.0 / banco de dados relacional.
C'_2= 1.0/ bancos de dados relacionais

3: Informações sobre banco de dados orientado a objetos.

C_3 = 1.0 / banco de dados orientado a objetos
C'_3 = 1.0 / bancos de dados orientados a objetos

4: Informações sobre banco de dados relacional difuso.

C_4 = 1.0 / banco de dados relacional difuso
C'_4 = 1.0 / bancos de dados difusos relacionais

5: Informações sobre banco de dados relacional, na preferência difuso.

C_8 = 1.0 / banco de dados relacional difuso + 0.5 / banco de dados relacional
C'_8 = 1.0 / bancos de dados relacionais difusos + 0.5 / bancos de dados relacionais

6: Informações sobre recuperação de informação.

C_7 = 1.0 / recuperação de informação
C'_7 = 1.0 / recuperando informações

7: Informações sobre recuperação de documentos.

C_7 = 1.0 / recuperação de documentos
C'_7 = 1.0 / recuperando documentos

8: Informações sobre recuperação de informação, na preferência recuperação de documentos.

C_8 = 1.0 / recuperação de documentos + 0.5 / recuperação de informação
C'_8 = 1.0 / recuperando documentos + 0.5 / recuperação de informações

ANEXO 6: Relevância dos documentos para as consultas

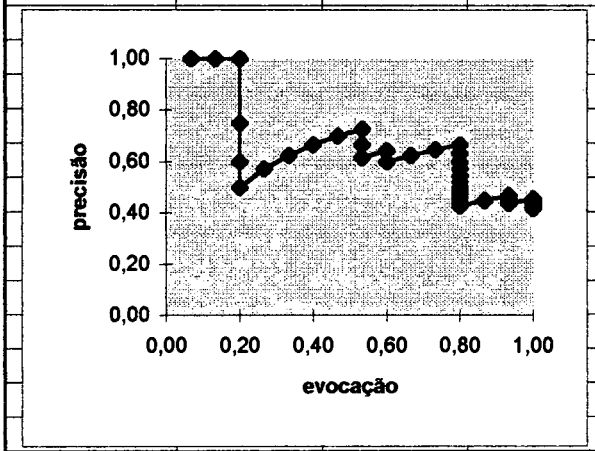
(+ = relevante; - = não relevante)

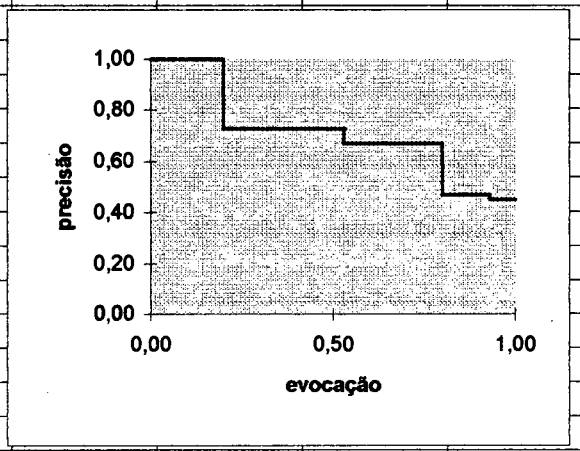
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
d_1	+	+	-	+	+	+	-	-
d_2	-	-	-	-	-	-	-	-
d_3	+	-	-	-	-	-	-	-
d_4	+	-	+	+	+	-	-	-
d_5	+	+	-	-	-	+	-	-
d_6	-	-	-	-	-	+	+	+
d_7	+	-	+	-	-	-	-	-
d_8	-	-	-	-	-	-	-	-
d_9	-	-	-	-	-	-	-	-
d_{10}	-	-	-	-	-	-	-	-
d_{11}	+	+	+	-	-	-	-	-
d_{12}	-	-	-	-	-	-	-	-
d_{13}	-	-	-	-	-	-	-	-
d_{14}	-	-	-	-	-	-	-	-
d_{15}	-	-	-	-	-	-	-	-
d_{16}	+	+	-	+	+	+	-	-
d_{17}	-	-	-	-	-	-	-	-
d_{18}	+	+	-	-	-	-	-	-
d_{19}	-	-	-	-	-	-	-	-
d_{20}	+	-	-	-	-	+	-	-
d_{21}	-	-	-	-	-	-	-	-
d_{22}	+	+	-	+	+	+	-	-
d_{23}	-	-	-	-	-	+	+	+
d_{24}	-	-	-	-	-	+	+	+
d_{25}	+	+	-	+	+	+	-	-
d_{26}	-	-	-	-	-	-	-	-
d_{27}	+	+	-	-	-	-	-	-
d_{28}	+	-	-	+	+	+	-	-
d_{29}	-	-	-	-	-	-	-	-
d_{30}	+	+	-	+	+	+	-	-
d_{31}	-	-	-	-	-	-	-	-
d_{32}	-	-	-	-	-	-	-	-
d_{33}	-	-	-	-	-	-	-	-
d_{34}	+	-	+	-	-	+	-	-
d_{35}	-	-	-	-	-	+	+	+
d_{36}	-	-	-	-	-	-	-	-
$\Sigma+$	15	9	4	7	7	13	4	4

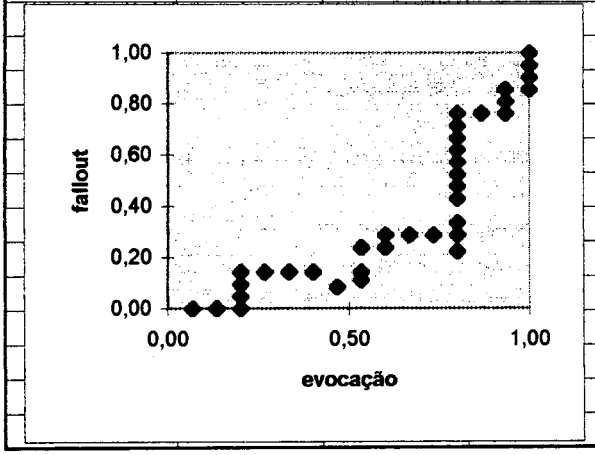
$\Sigma+$ = soma dos documentos relevantes para a consulta

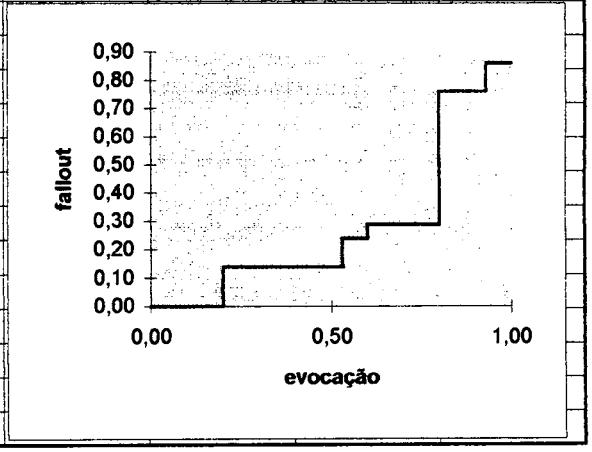
Anexo 7: Exemplos de respostas e determinação da precisão e do fallout nos níveis da evocação

Miyamoto							
consulta:	1		total rel.	15			
n	no doc	relevante	rel. ac.	relevância	evocação	precisão	fallout
1	3	1	1	1,00	0,07	1,00	0,00
2	7	1	2	1,00	0,13	1,00	0,00
3	20	1	3	1,00	0,20	1,00	0,00
4	2	0	3	0,33	0,20	0,75	0,05
5	8	0	3	0,33	0,20	0,60	0,10
6	32	0	3	0,33	0,20	0,50	0,14
7	4	1	4	0,25	0,27	0,57	0,14
8	11	1	5	0,25	0,33	0,63	0,14
9	27	1	6	0,25	0,40	0,67	0,14
10	29	1	7	0,25	0,47	0,70	0,08
11	34	1	8	0,25	0,53	0,73	0,14
12	1	0	8	0,16	0,53	0,67	0,11
13	14	0	8	0,16	0,53	0,62	0,24
14	16	1	9	0,16	0,60	0,64	0,24
15	31	0	9	0,16	0,60	0,60	0,29
16	5	1	10	0,14	0,67	0,63	0,29
17	25	1	11	0,14	0,73	0,65	0,29
18	28	1	12	0,14	0,80	0,67	0,29
19	10	0	12	0,09	0,80	0,63	0,33
20	13	0	12	0,09	0,80	0,60	0,22
21	17	0	12	0,09	0,80	0,57	0,43
22	19	0	12	0,09	0,80	0,55	0,48
23	23	0	12	0,09	0,80	0,52	0,52
24	36	0	12	0,09	0,80	0,50	0,57
25	6	0	12	0,09	0,80	0,48	0,62
26	15	0	12	0,07	0,80	0,46	0,67
27	24	0	12	0,07	0,80	0,44	0,71
28	35	0	12	0,07	0,80	0,43	0,76
29	22	1	13	0,07	0,87	0,45	0,76
30	30	1	14	0,07	0,93	0,47	0,76
31	9	0	14	0,00	0,93	0,45	0,81
32	12	0	14	0,00	0,93	0,44	0,86
33	18	1	15	0,00	1,00	0,45	0,86
34	21	0	15	0,00	1,00	0,44	0,90
35	26	0	15	0,00	1,00	0,43	0,95
36	33	0	15	0,00	1,00	0,42	1,00

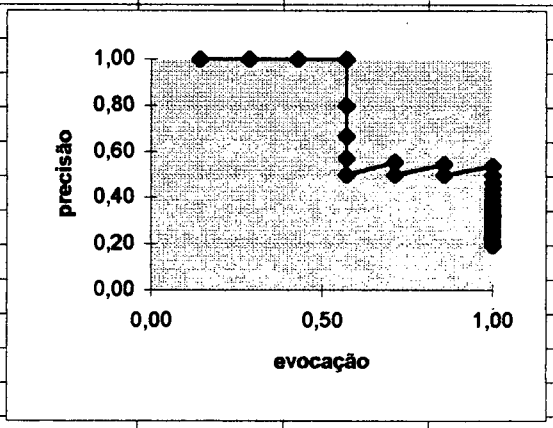


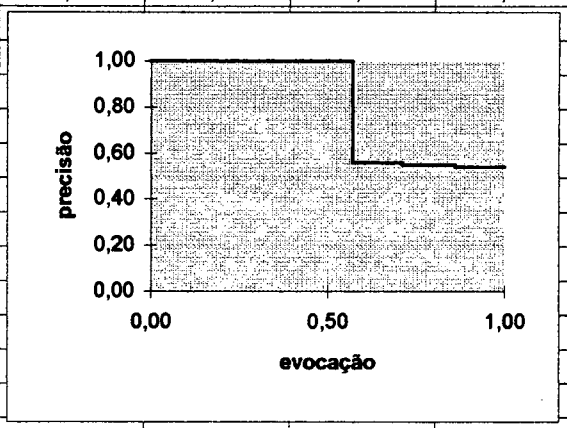


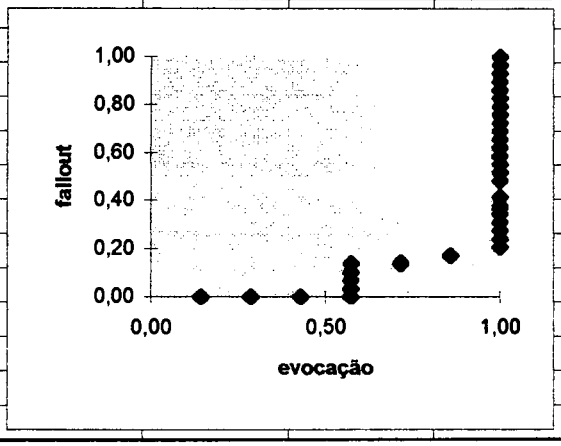


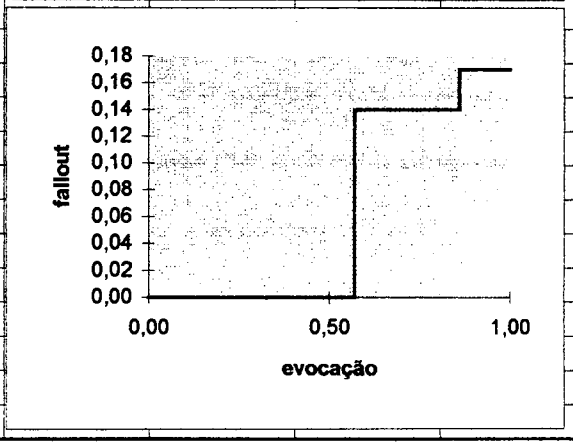


Radix	dic. vazio	descritores controlados			consultas controladas		
consulta:	4		total rel.	7			
n	no doc	relevante	rel. ac.	relevância	evocação	precisão	fallout
1	1	1	1	1,00	0,14	1,00	0,00
2	16	1	2	1,00	0,29	1,00	0,00
3	22	1	3	1,00	0,43	1,00	0,00
4	30	1	4	1,00	0,57	1,00	0,00
5	3	0	4	0,80	0,57	0,80	0,03
6	5	0	4	0,80	0,57	0,67	0,07
7	7	0	4	0,80	0,57	0,57	0,10
8	20	0	4	0,80	0,57	0,50	0,14
9	25	1	5	0,80	0,71	0,56	0,14
10	27	0	5	0,80	0,71	0,50	0,14
11	28	1	6	0,80	0,86	0,55	0,17
12	34	0	6	0,80	0,86	0,50	0,17
13	4	1	7	0,78	1,00	0,54	0,21
14	11	0	7	0,78	1,00	0,50	0,24
15	9	0	7	0,30	1,00	0,47	0,28
16	26	0	7	0,30	1,00	0,44	0,31
17	6	0	7	0,29	1,00	0,41	0,34
18	23	0	7	0,29	1,00	0,39	0,38
19	2	0	7	0,23	1,00	0,37	0,41
20	8	0	7	0,23	1,00	0,35	0,37
21	12	0	7	0,23	1,00	0,33	0,48
22	18	0	7	0,23	1,00	0,32	0,52
23	32	0	7	0,23	1,00	0,30	0,55
24	21	0	7	0,22	1,00	0,29	0,59
25	15	0	7	0,21	1,00	0,28	0,62
26	33	0	7	0,21	1,00	0,27	0,66
27	36	0	7	0,21	1,00	0,26	0,69
28	13	0	7	0,19	1,00	0,25	0,72
29	10	0	7	0,18	1,00	0,24	0,76
30	14	0	7	0,18	1,00	0,23	0,79
31	17	0	7	0,18	1,00	0,23	0,83
32	19	0	7	0,18	1,00	0,22	0,86
33	24	0	7	0,18	1,00	0,21	0,90
34	29	0	7	0,18	1,00	0,21	0,93
35	31	0	7	0,18	1,00	0,20	0,97
36	35	0	7	0,17	1,00	0,19	1,00

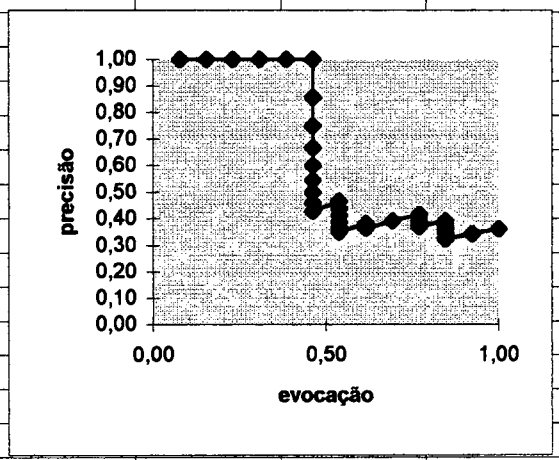


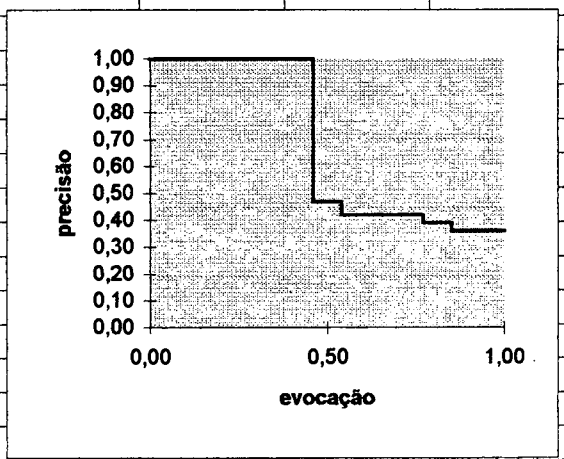


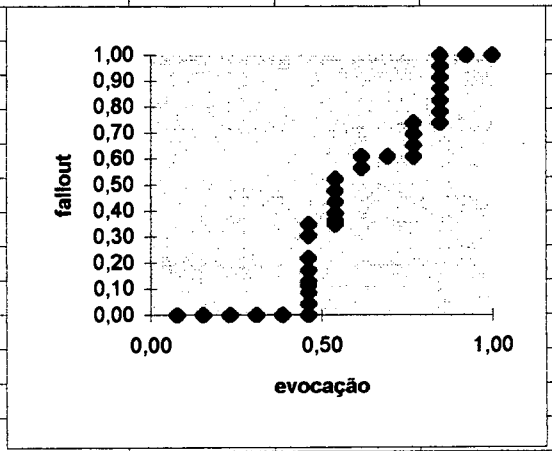


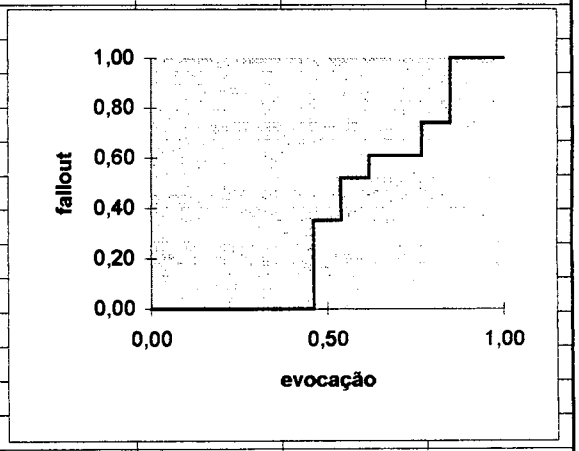


Radix	dic. não-vazio	descritores controlados			consultas controladas		
consulta:	6		total rel.	13			
n	no doc	relevante	rel. ac.	relevância	evocação	precisão	fallout
1	1	1	1	1,00	0,08	1,00	0,00
2	16	1	2	1,00	0,15	1,00	0,00
3	24	1	3	1,00	0,23	1,00	0,00
4	35	1	4	1,00	0,31	1,00	0,00
5	6	1	5	0,80	0,38	1,00	0,00
6	23	1	6	0,80	0,46	1,00	0,00
7	18	0	6	0,74	0,46	0,86	0,04
8	36	0	6	0,71	0,46	0,75	0,09
9	4	0	6	0,69	0,46	0,67	0,13
10	15	0	6	0,69	0,46	0,60	0,11
11	19	0	6	0,69	0,46	0,55	0,22
12	27	0	6	0,29	0,46	0,50	0,17
13	32	0	6	0,29	0,46	0,46	0,30
14	33	0	6	0,29	0,46	0,43	0,35
15	5	1	7	0,21	0,54	0,47	0,35
16	8	0	7	0,21	0,54	0,44	0,39
17	11	0	7	0,21	0,54	0,41	0,43
18	12	0	7	0,21	0,54	0,39	0,48
19	13	0	7	0,21	0,54	0,37	0,52
20	21	0	7	0,21	0,54	0,35	0,36
21	25	1	8	0,21	0,62	0,38	0,57
22	26	0	8	0,21	0,62	0,36	0,61
23	28	1	9	0,21	0,69	0,39	0,61
24	34	1	10	0,21	0,77	0,42	0,61
25	10	0	10	0,20	0,77	0,40	0,65
26	9	0	10	0,19	0,77	0,38	0,70
27	17	0	10	0,19	0,77	0,37	0,74
28	20	1	11	0,19	0,85	0,39	0,74
29	29	0	11	0,19	0,85	0,38	0,78
30	31	0	11	0,19	0,85	0,37	0,83
31	2	0	11	0,18	0,85	0,35	0,87
32	14	0	11	0,18	0,85	0,34	0,91
33	3	0	11	0,17	0,85	0,33	0,96
34	7	0	11	0,17	0,85	0,32	1,00
35	22	1	12	0,17	0,92	0,34	1,00
36	30	1	13	0,17	1,00	0,36	1,00

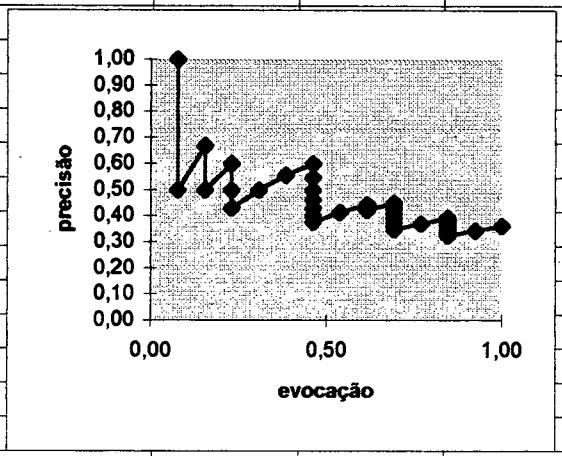


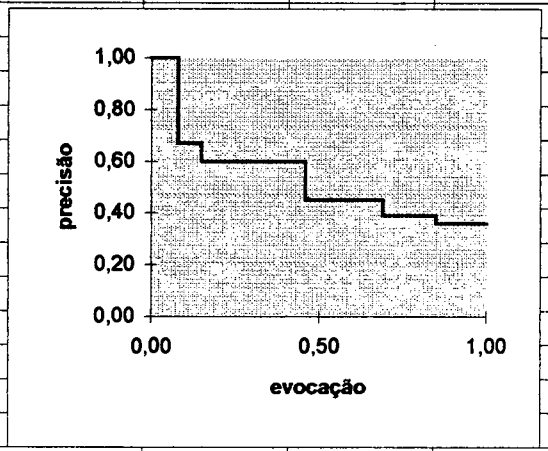


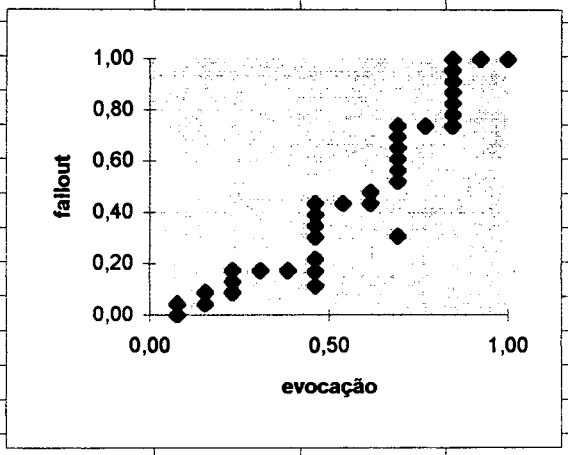


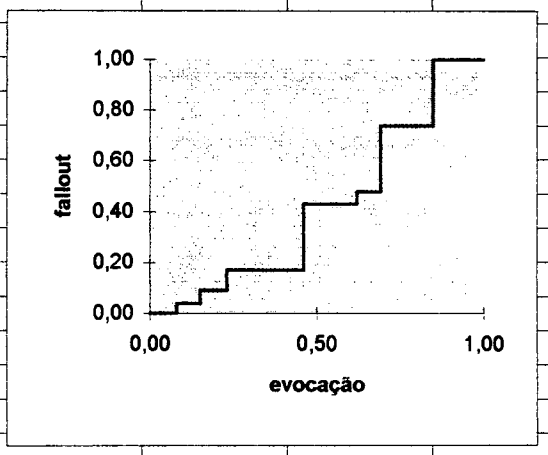


Radlx	dic. vazio		descritores controlados			consultas não-controladas	
consulta:	6		total rel.	13			
n	no doc	relevante	rel. ac.	relevância	evocação	precisão	fallout
1	1	1	1	0,62	0,08	1,00	0,00
2	4	0	1	0,62	0,08	0,50	0,04
3	6	1	2	0,62	0,15	0,67	0,04
4	15	0	2	0,62	0,15	0,50	0,09
5	16	1	3	0,62	0,23	0,60	0,09
6	18	0	3	0,62	0,23	0,50	0,13
7	19	0	3	0,62	0,23	0,43	0,17
8	23	1	4	0,62	0,31	0,50	0,17
9	24	1	5	0,62	0,38	0,56	0,17
10	35	1	6	0,62	0,46	0,60	0,11
11	36	0	6	0,62	0,46	0,55	0,22
12	14	0	6	0,28	0,46	0,50	0,17
13	27	0	6	0,28	0,46	0,46	0,30
14	32	0	6	0,28	0,46	0,43	0,35
15	33	0	6	0,28	0,46	0,40	0,39
16	2	0	6	0,25	0,46	0,38	0,43
17	20	1	7	0,25	0,54	0,41	0,43
18	25	1	8	0,25	0,62	0,44	0,43
19	29	0	8	0,25	0,62	0,42	0,48
20	5	1	9	0,21	0,69	0,45	0,31
21	8	0	9	0,21	0,69	0,43	0,52
22	11	0	9	0,21	0,69	0,41	0,57
23	12	0	9	0,21	0,69	0,39	0,61
24	13	0	9	0,21	0,69	0,38	0,65
25	21	0	9	0,21	0,69	0,36	0,70
26	26	0	9	0,21	0,69	0,35	0,74
27	28	1	10	0,21	0,77	0,37	0,74
28	34	1	11	0,21	0,85	0,39	0,74
29	10	0	11	0,19	0,85	0,38	0,78
30	9	0	11	0,19	0,85	0,37	0,83
31	17	0	11	0,19	0,85	0,35	0,87
32	31	0	11	0,19	0,85	0,34	0,91
33	3	0	11	0,17	0,85	0,33	0,96
34	7	0	11	0,17	0,85	0,32	1,00
35	22	1	12	0,17	0,92	0,34	1,00
36	30	1	13	0,17	1,00	0,36	1,00

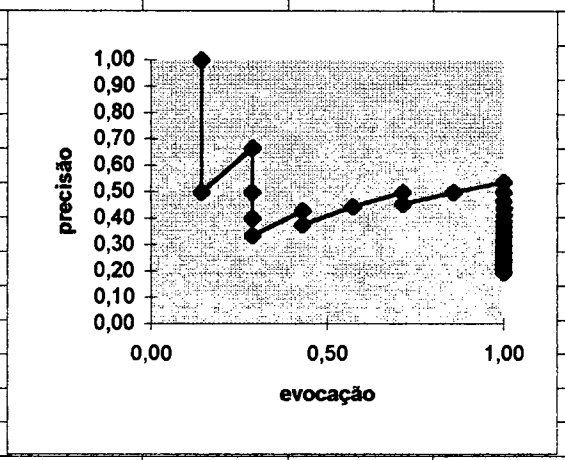


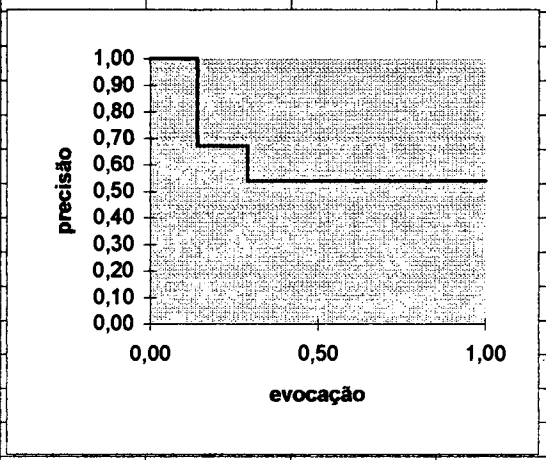


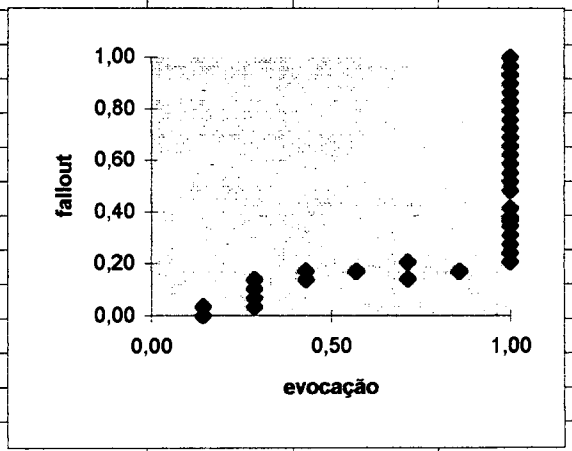


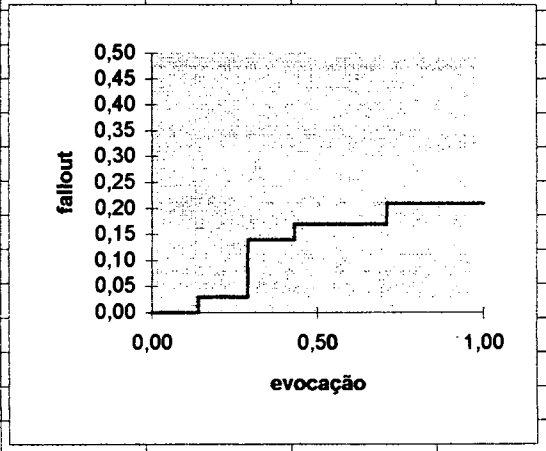


Radix	dic. não-vazio	descritores controlados	consultas não-controladas				
consulta:	4		total rel.	7			
n	no doc	relevante	rel. ac.	relevância	evocação	precisão	fallout
1	1	1	1	0,55	0,14	1,00	0,00
2	3	0	1	0,55	0,14	0,50	0,03
3	4	1	2	0,55	0,29	0,67	0,03
4	5	0	2	0,55	0,29	0,50	0,07
5	7	0	2	0,55	0,29	0,40	0,10
6	11	0	2	0,55	0,29	0,33	0,14
7	16	1	3	0,55	0,43	0,43	0,14
8	20	0	3	0,55	0,43	0,38	0,17
9	22	1	4	0,55	0,57	0,44	0,17
10	25	1	5	0,55	0,71	0,50	0,14
11	27	0	5	0,55	0,71	0,45	0,21
12	28	1	6	0,55	0,86	0,50	0,17
13	30	1	7	0,55	1,00	0,54	0,21
14	34	0	7	0,55	1,00	0,50	0,24
15	9	0	7	0,30	1,00	0,47	0,28
16	26	0	7	0,30	1,00	0,44	0,31
17	6	0	7	0,29	1,00	0,41	0,34
18	23	0	7	0,29	1,00	0,39	0,38
19	2	0	7	0,23	1,00	0,37	0,41
20	8	0	7	0,23	1,00	0,35	0,37
21	12	0	7	0,23	1,00	0,33	0,48
22	18	0	7	0,23	1,00	0,32	0,52
23	32	0	7	0,23	1,00	0,30	0,55
24	21	0	7	0,22	1,00	0,29	0,59
25	15	0	7	0,21	1,00	0,28	0,62
26	33	0	7	0,21	1,00	0,27	0,66
27	36	0	7	0,21	1,00	0,26	0,69
28	13	0	7	0,19	1,00	0,25	0,72
29	10	0	7	0,18	1,00	0,24	0,76
30	14	0	7	0,18	1,00	0,23	0,79
31	17	0	7	0,18	1,00	0,23	0,83
32	19	0	7	0,18	1,00	0,22	0,86
33	24	0	7	0,18	1,00	0,21	0,90
34	29	0	7	0,18	1,00	0,21	0,93
35	31	0	7	0,18	1,00	0,20	0,97
36	35	0	7	0,17	1,00	0,19	1,00

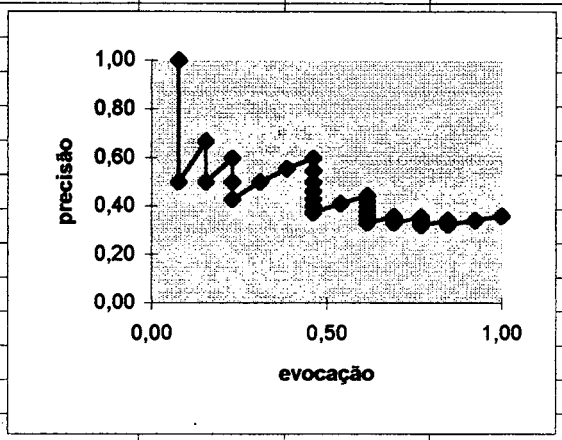


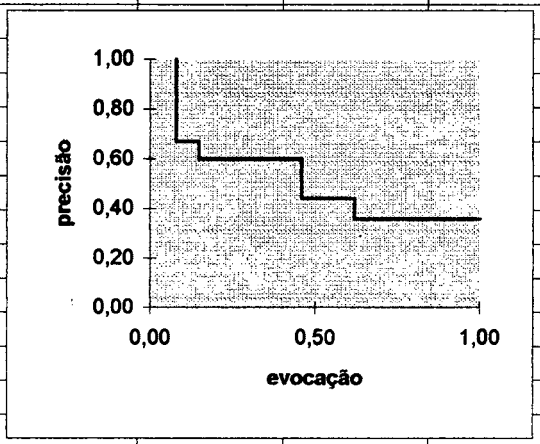


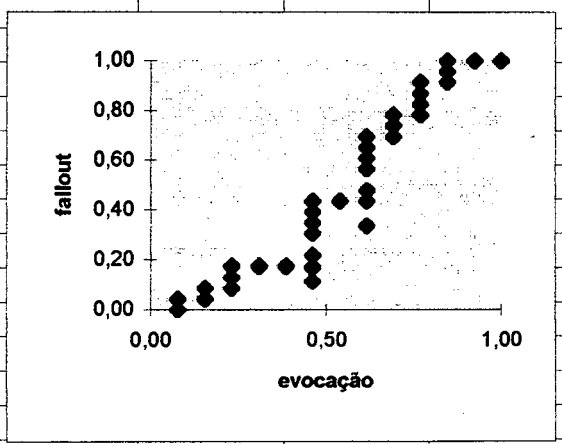


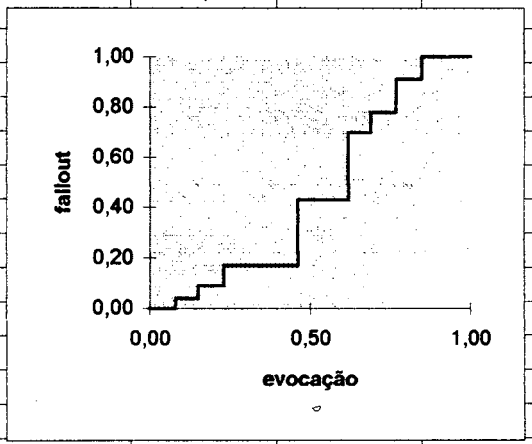


Radix	dic. vazio	descritores não-controlados			consultas não-controladas		
consulta:	6		total rel.	13			
n	no doc	relevante	rel. ac.	relevância	evocação	precisão	fallout
1	1	1	1	0,68	0,08	1,00	0,00
2	4	0	1	0,68	0,08	0,50	0,04
3	6	1	2	0,68	0,15	0,67	0,04
4	15	0	2	0,68	0,15	0,50	0,09
5	16	1	3	0,68	0,23	0,60	0,09
6	18	0	3	0,68	0,23	0,50	0,13
7	19	0	3	0,68	0,23	0,43	0,17
8	23	1	4	0,68	0,31	0,50	0,17
9	24	1	5	0,68	0,38	0,56	0,17
10	35	1	6	0,68	0,46	0,60	0,11
11	36	0	6	0,67	0,46	0,55	0,22
12	14	0	6	0,28	0,46	0,50	0,17
13	27	0	6	0,28	0,46	0,46	0,30
14	32	0	6	0,28	0,46	0,43	0,35
15	33	0	6	0,28	0,46	0,40	0,39
16	2	0	6	0,25	0,46	0,38	0,43
17	20	1	7	0,25	0,54	0,41	0,43
18	25	1	8	0,25	0,62	0,44	0,43
19	29	0	8	0,25	0,62	0,42	0,48
20	9	0	8	0,22	0,62	0,40	0,34
21	10	0	8	0,22	0,62	0,38	0,57
22	12	0	8	0,22	0,62	0,36	0,61
23	13	0	8	0,22	0,62	0,35	0,65
24	17	0	8	0,22	0,62	0,33	0,70
25	28	1	9	0,22	0,69	0,36	0,70
26	31	0	9	0,22	0,69	0,35	0,74
27	8	0	9	0,19	0,69	0,33	0,78
28	5	1	10	0,19	0,77	0,36	0,78
29	11	0	10	0,19	0,77	0,34	0,83
30	21	0	10	0,19	0,77	0,33	0,87
31	26	0	10	0,19	0,77	0,32	0,91
32	34	1	11	0,19	0,85	0,34	0,91
33	3	0	11	0,18	0,85	0,33	0,96
34	7	0	11	0,18	0,85	0,32	1,00
35	22	1	12	0,17	0,92	0,34	1,00
36	30	1	13	0,17	1,00	0,36	1,00

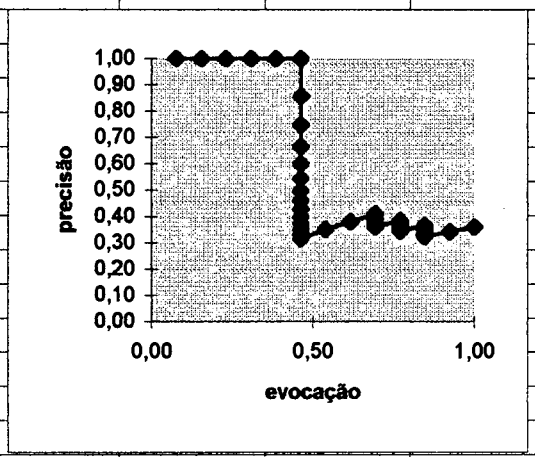


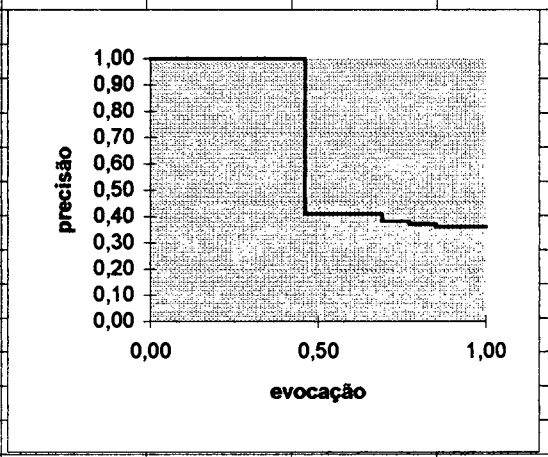


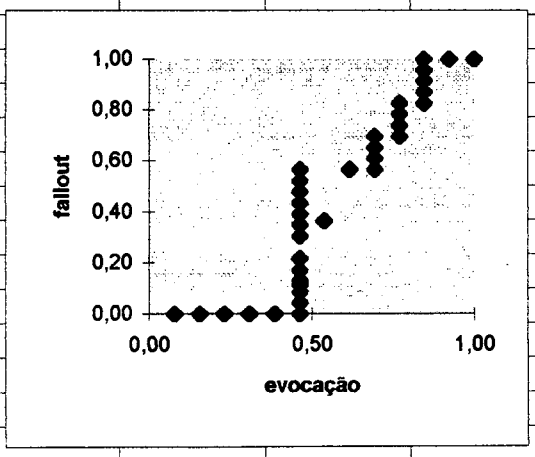


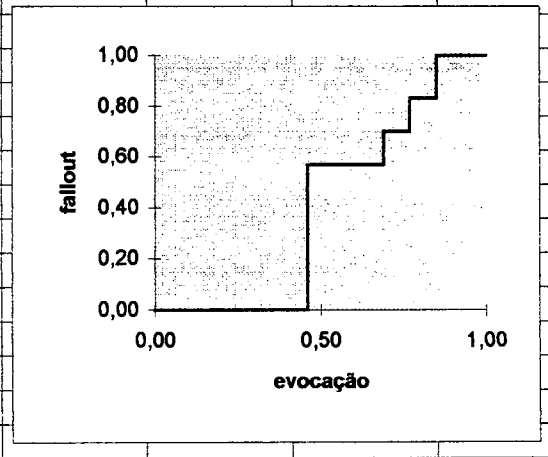


Radix	dic. não-vazio	descritores não-controlados			consultas não-controladas		
consulta:	6		total rel.	13			
n	no doc	relevante	rel. ac.	relevância	evocação	precisão	fallout
1	24	1	1	1,00	0,08	1,00	0,00
2	1	1	2	0,90	0,15	1,00	0,00
3	16	1	3	0,90	0,23	1,00	0,00
4	35	1	4	0,90	0,31	1,00	0,00
5	6	1	5	0,84	0,38	1,00	0,00
6	23	1	6	0,84	0,46	1,00	0,00
7	15	0	6	0,73	0,46	0,86	0,04
8	18	0	6	0,73	0,46	0,75	0,09
9	19	0	6	0,73	0,46	0,67	0,13
10	36	0	6	0,72	0,46	0,60	0,11
11	4	0	6	0,70	0,46	0,55	0,22
12	27	0	6	0,30	0,46	0,50	0,17
13	32	0	6	0,30	0,46	0,46	0,30
14	33	0	6	0,30	0,46	0,43	0,35
15	9	0	6	0,22	0,46	0,40	0,39
16	10	0	6	0,22	0,46	0,38	0,43
17	12	0	6	0,22	0,46	0,35	0,48
18	13	0	6	0,22	0,46	0,33	0,52
19	17	0	6	0,22	0,46	0,32	0,57
20	20	1	7	0,22	0,54	0,35	0,36
21	25	1	8	0,22	0,62	0,38	0,57
22	28	1	9	0,22	0,69	0,41	0,57
23	29	0	9	0,22	0,69	0,39	0,61
24	31	0	9	0,22	0,69	0,38	0,65
25	8	0	9	0,19	0,69	0,36	0,70
26	5	1	10	0,19	0,77	0,38	0,70
27	11	0	10	0,19	0,77	0,37	0,74
28	21	0	10	0,19	0,77	0,36	0,78
29	26	0	10	0,19	0,77	0,34	0,83
30	34	1	11	0,19	0,85	0,37	0,83
31	2	0	11	0,18	0,85	0,35	0,87
32	14	0	11	0,18	0,85	0,34	0,91
33	3	0	11	0,18	0,85	0,33	0,96
34	7	0	11	0,18	0,85	0,32	1,00
35	22	1	12	0,17	0,92	0,34	1,00
36	30	1	13	0,17	1,00	0,36	1,00









Anexo 8: Médias de evocação, precisão e fallout nos níveis do threshold e gráficos de precisão/evocação e de fallout/evocação

Miyamoto

Médias nos níveis do threshold (alpha)

evocação

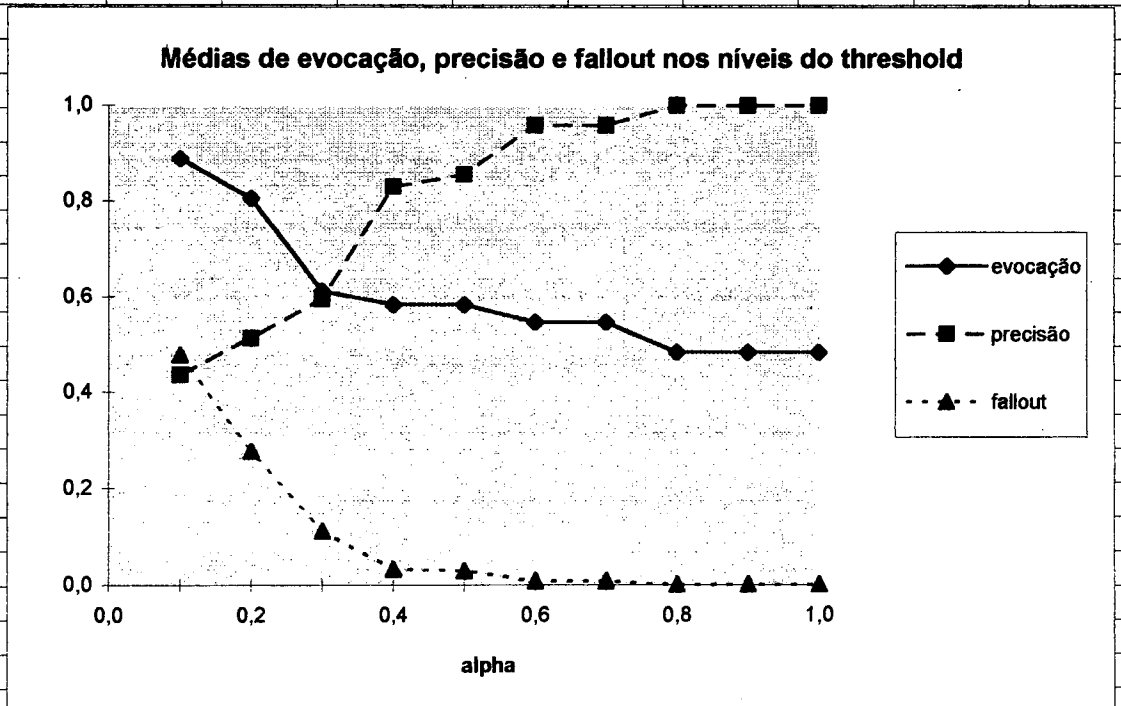
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,80	0,89	1,00	1,00	1,00	0,92	0,50	1,00	0,889
0.2	0,53	0,78	1,00	0,86	1,00	0,77	0,50	1,00	0,805
0.3	0,20	0,22	1,00	0,57	0,86	0,54	0,50	1,00	0,611
0.4	0,20	0,22	1,00	0,57	0,86	0,31	0,50	1,00	0,583
0.5	0,20	0,22	1,00	0,57	0,86	0,31	0,50	1,00	0,583
0.6	0,20	0,22	1,00	0,57	0,57	0,31	0,50	1,00	0,546
0.7	0,20	0,22	1,00	0,57	0,57	0,31	0,50	1,00	0,546
0.8	0,20	0,22	1,00	0,57	0,57	0,31	0,50	0,50	0,484
0.9	0,20	0,22	1,00	0,57	0,57	0,31	0,50	0,50	0,484
1.0	0,20	0,22	1,00	0,57	0,57	0,31	0,50	0,50	0,484

precisao

alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,67	0,50	0,17	0,27	0,27	0,46	1,00	0,15	0,436
0.2	0,73	0,54	0,22	0,40	0,41	0,59	1,00	0,21	0,513
0.3	0,50	0,50	0,67	0,50	0,55	0,70	1,00	0,33	0,594
0.4	1,00	0,50	0,80	1,00	0,67	1,00	1,00	0,67	0,830
0.5	1,00	0,50	1,00	1,00	0,67	1,00	1,00	0,67	0,855
0.6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,67	0,959
0.7	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,67	0,959
0.8	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
1.0	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000

fallout

alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,29	0,30	0,59	0,66	0,66	0,61	0,00	0,72	0,479
0.2	0,14	0,22	0,44	0,31	0,34	0,30	0,00	0,47	0,278
0.3	0,14	0,07	0,06	0,14	0,17	0,08	0,00	0,23	0,111
0.4	0,00	0,07	0,03	0,00	0,10	0,00	0,00	0,06	0,033
0.5	0,00	0,07	0,00	0,00	0,10	0,00	0,00	0,06	0,029
0.6	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,008
0.7	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,008
0.8	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000



Miyamoto

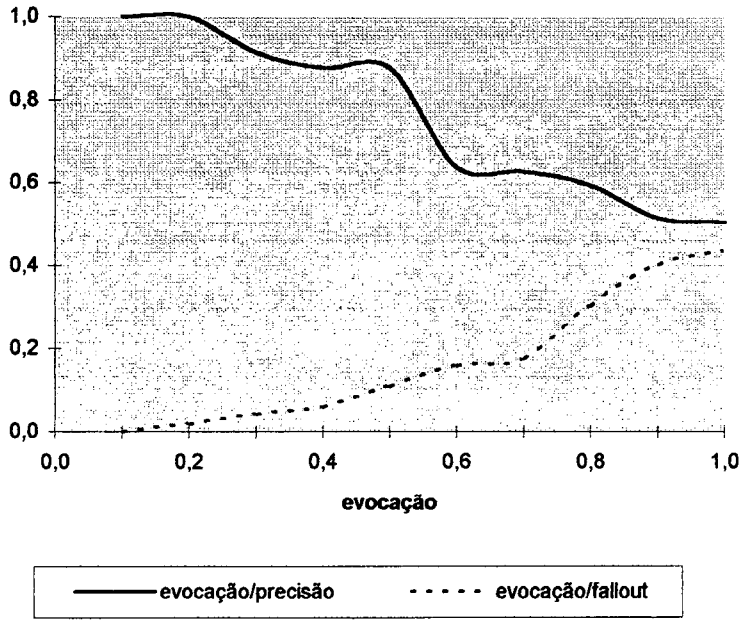
Médias da precisão nos níveis da evocação

evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.3	0,73	0,58	1,00	1,00	1,00	1,00	1,00	1,00	0,914
0.4	0,73	0,58	1,00	1,00	1,00	0,70	1,00	1,00	0,876
0.5	0,73	0,58	1,00	1,00	1,00	0,70	1,00	1,00	0,876
0.6	0,67	0,58	1,00	0,56	0,75	0,67	0,21	0,67	0,639
0.7	0,67	0,58	1,00	0,56	0,75	0,59	0,21	0,67	0,629
0.8	0,67	0,50	1,00	0,43	0,75	0,52	0,21	0,67	0,594
0.9	0,47	0,26	1,00	0,41	0,58	0,52	0,21	0,67	0,515
1.0	0,45	0,26	1,00	0,41	0,58	0,46	0,21	0,67	0,505

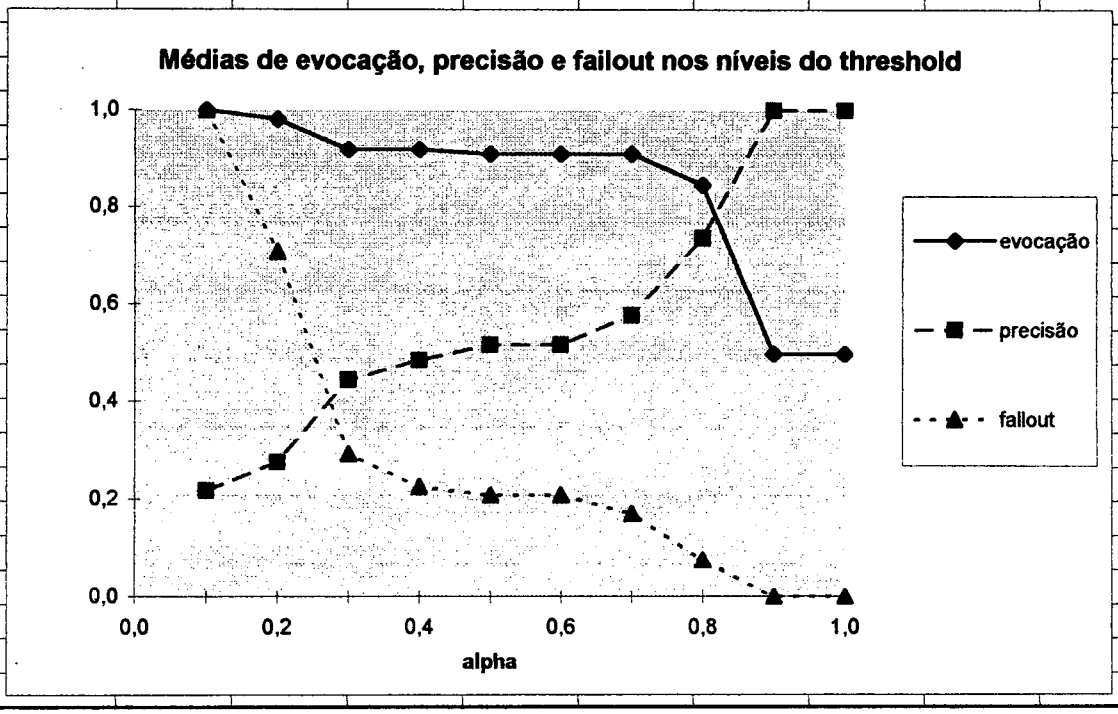
Médias do fallout nos níveis da evocação

evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.2	0,14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,018
0.3	0,14	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,041
0.4	0,14	0,19	0,00	0,00	0,00	0,13	0,00	0,00	0,058
0.5	0,14	0,19	0,00	0,00	0,00	0,13	0,34	0,06	0,108
0.6	0,29	0,19	0,00	0,14	0,07	0,17	0,34	0,06	0,158
0.7	0,29	0,19	0,00	0,14	0,07	0,30	0,34	0,06	0,174
0.8	0,76	0,30	0,00	0,28	0,07	0,48	0,47	0,06	0,303
0.9	0,76	0,93	0,00	0,34	0,17	0,48	0,47	0,06	0,401
1.0	0,86	0,93	0,00	0,34	0,17	0,65	0,47	0,06	0,435

Gráficos de evocação/precisão e evocação/fallout em média



Radix			consultas controladas						
dicionário vazio			descritores controlados						
		Médias nos níveis do threshold (alpha)							
evocação									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	0,85	1,00	1,00	0,981
0.3	1,00	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,919
0.4	1,00	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,919
0.5	0,93	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,910
0.6	0,93	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,910
0.7	0,93	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,910
0.8	0,80	0,78	1,00	0,86	0,86	0,46	1,00	1,00	0,845
0.9	0,20	0,33	1,00	0,57	0,57	0,31	0,50	0,50	0,498
1.0	0,20	0,33	1,00	0,57	0,57	0,31	0,50	0,50	0,498
precisao									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,42	0,25	0,11	0,19	0,19	0,36	0,11	0,11	0,218
0.2	0,54	0,32	0,15	0,26	0,26	0,39	0,14	0,14	0,275
0.3	0,65	0,50	0,25	0,44	0,44	0,55	0,36	0,36	0,444
0.4	0,75	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,485
0.5	1,00	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,516
0.6	1,00	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,516
0.7	1,00	0,57	0,29	0,50	0,50	0,75	0,50	0,50	0,576
0.8	1,00	0,55	1,00	0,50	0,50	1,00	0,67	0,67	0,736
0.9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
1.0	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
fallout									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	0,62	0,70	0,72	0,69	0,69	0,74	0,75	0,75	0,708
0.3	0,38	0,30	0,38	0,31	0,31	0,22	0,22	0,22	0,293
0.4	0,14	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,226
0.5	0,00	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,209
0.6	0,00	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,209
0.7	0,00	0,22	0,31	0,24	0,24	0,09	-0,13	0,13	0,170
0.8	0,00	0,14	0,00	0,17	0,17	0,00	0,06	0,06	0,075
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000



Radix			consultas controladas					
dicionário vazio			descritores controlados					

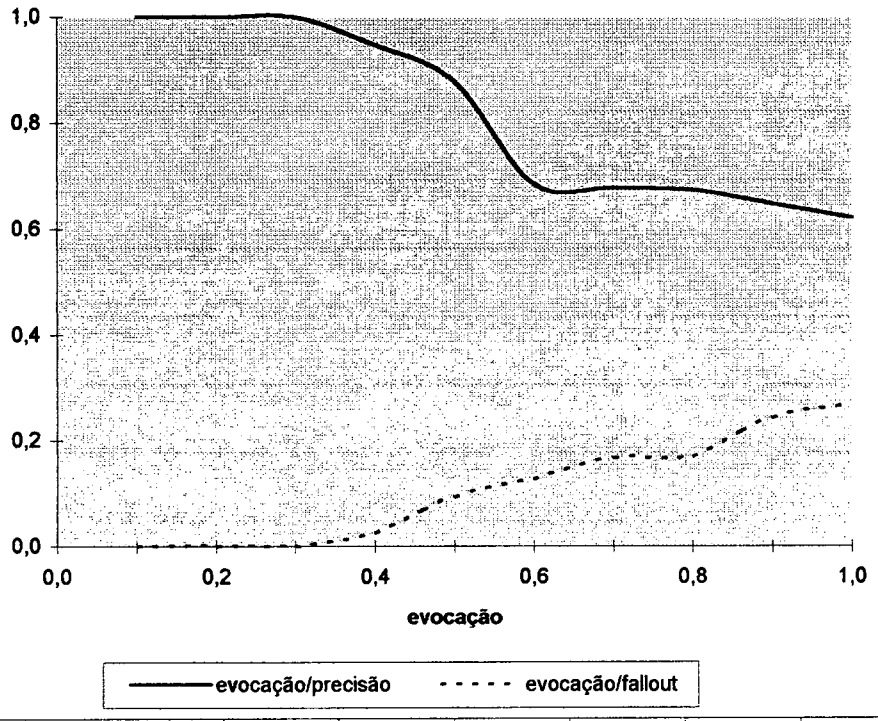
Médias da precisão nos níveis da evocação

evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.3	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.4	1,00	0,58	1,00	1,00	1,00	1,00	1,00	1,00	0,948
0.5	1,00	0,58	1,00	1,00	1,00	0,45	1,00	1,00	0,879
0.6	1,00	0,58	1,00	0,56	0,56	0,45	0,67	0,67	0,686
0.7	1,00	0,58	1,00	0,56	0,56	0,39	0,67	0,67	0,679
0.8	1,00	0,57	1,00	0,55	0,55	0,39	0,67	0,67	0,675
0.9	1,00	0,41	1,00	0,54	0,54	0,36	0,67	0,67	0,649
1.0	0,79	0,41	1,00	0,54	0,54	0,36	0,67	0,67	0,623

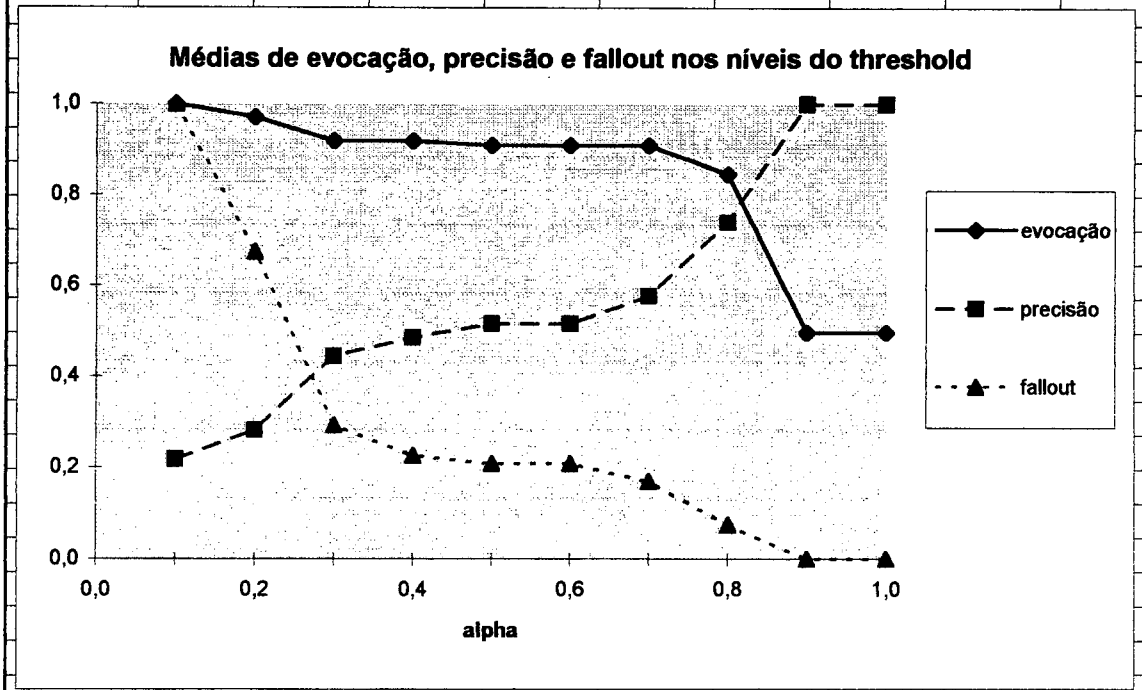
Médias do fallout nos níveis da evocação

evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.2	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.3	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.4	0,00	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,024
0.5	0,00	0,19	0,00	0,00	0,00	0,43	0,06	0,06	0,093
0.6	0,00	0,19	0,00	0,14	0,14	0,43	0,06	0,06	0,128
0.7	0,00	0,19	0,00	0,14	0,14	0,74	0,06	0,06	0,166
0.8	0,00	0,22	0,00	0,14	0,14	0,74	0,06	0,06	0,170
0.9	0,00	0,48	0,00	0,17	0,17	1,00	0,06	0,06	0,243
1.0	0,19	0,48	0,00	0,17	0,17	1,00	0,06	0,06	0,266

Gráficos de evocação/precisão e evocação/fallout em média



Radix			consultas controladas						
dicionário não-vazio			descritores controlados						
		Médias nos níveis do threshold (alpha)							
evocação									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	0,77	1,00	1,00	0,971
0.3	1,00	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,919
0.4	1,00	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,919
0.5	0,93	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,910
0.6	0,93	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,910
0.7	0,93	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,910
0.8	0,80	0,78	1,00	0,86	0,86	0,46	1,00	1,00	0,845
0.9	0,20	0,33	1,00	0,57	0,57	0,31	0,50	0,50	0,498
1.0	0,20	0,33	1,00	0,57	0,57	0,31	0,50	0,50	0,498
precisao									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,42	0,25	0,11	0,19	0,19	0,36	0,11	0,11	0,218
0.2	0,54	0,32	0,15	0,26	0,26	0,40	0,16	0,16	0,281
0.3	0,65	0,50	0,25	0,44	0,44	0,55	0,36	0,36	0,444
0.4	0,75	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,485
0.5	1,00	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,516
0.6	1,00	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,516
0.7	1,00	0,57	0,29	0,50	0,50	0,75	0,50	0,50	0,576
0.8	1,00	0,58	1,00	0,50	0,50	1,00	0,67	0,67	0,740
0.9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
1.0	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
fallout									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	0,62	0,70	0,72	0,69	0,69	0,65	0,66	0,66	0,674
0.3	0,38	0,30	0,38	0,31	0,31	0,22	0,22	0,22	0,293
0.4	0,14	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,226
0.5	0,00	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,209
0.6	0,00	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,209
0.7	0,00	0,22	0,31	0,24	0,24	0,09	0,13	0,13	0,170
0.8	0,00	0,14	0,00	0,17	0,17	0,00	0,06	0,06	0,075
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000



Radix			consultas controladas					
dicionário não-vazio			descritores controlados					

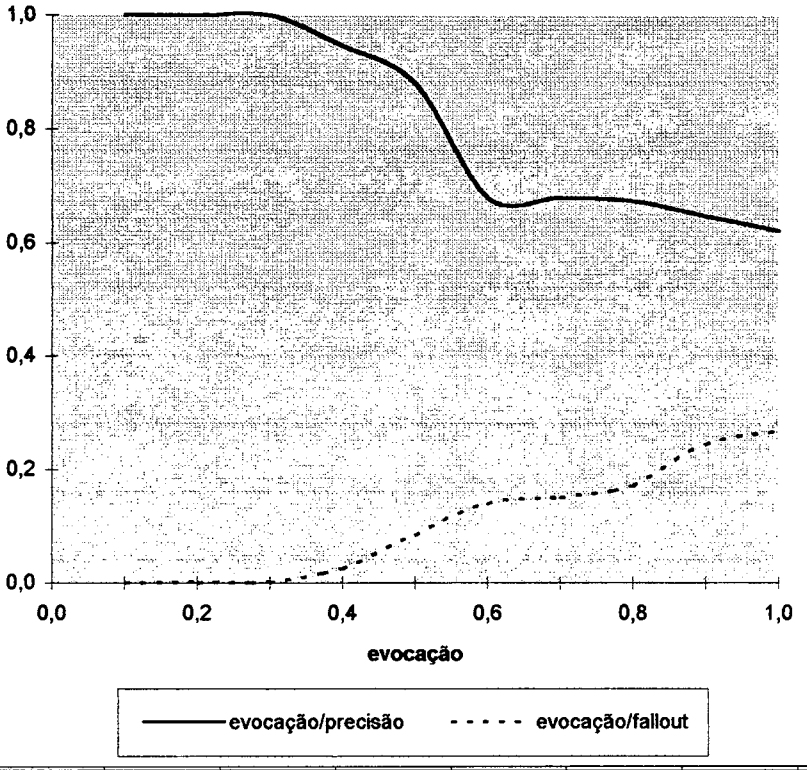
Médias da precisão nos níveis da evocação

evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.3	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.4	1,00	0,57	1,00	1,00	1,00	1,00	1,00	1,00	0,946
0.5	1,00	0,57	1,00	1,00	1,00	0,47	1,00	1,00	0,880
0.6	1,00	0,57	1,00	0,56	0,56	0,42	0,67	0,67	0,681
0.7	1,00	0,57	1,00	0,56	0,56	0,42	0,67	0,67	0,681
0.8	1,00	0,57	1,00	0,55	0,55	0,39	0,67	0,67	0,675
0.9	1,00	0,41	1,00	0,54	0,54	0,36	0,67	0,67	0,649
1.0	0,79	0,41	1,00	0,54	0,54	0,36	0,67	0,67	0,623

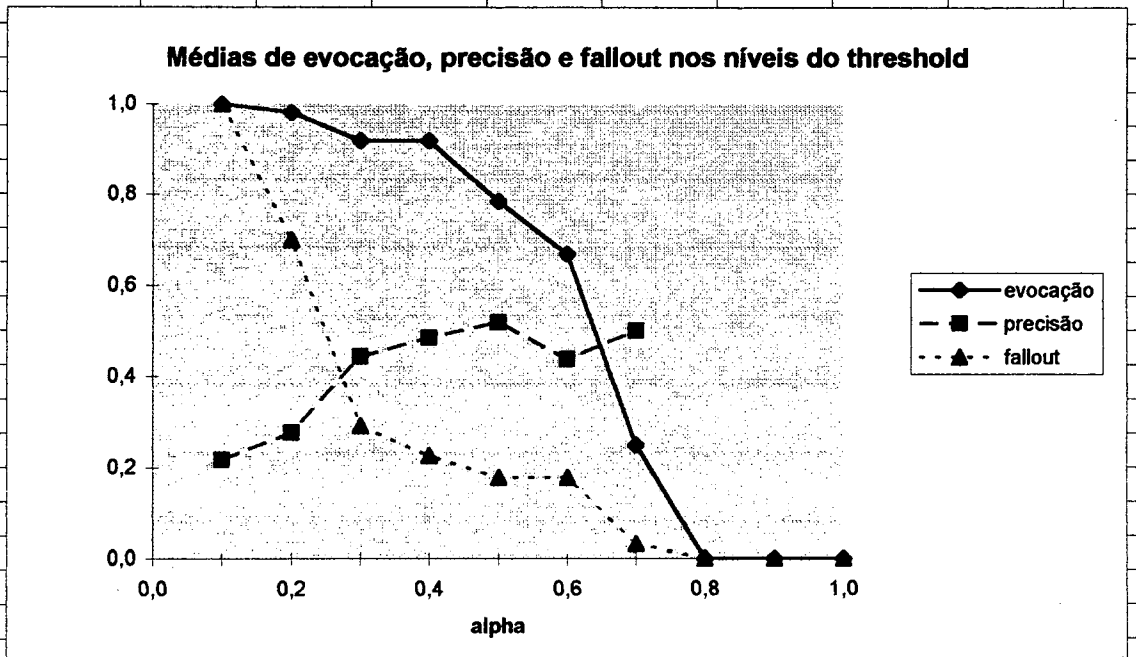
Médias do fallout nos níveis da evocação

evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.2	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.3	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.4	0,00	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,024
0.5	0,00	0,19	0,00	0,00	0,00	0,35	0,06	0,06	0,083
0.6	0,00	0,19	0,00	0,14	0,14	0,52	0,06	0,06	0,139
0.7	0,00	0,19	0,00	0,14	0,14	0,61	0,06	0,06	0,150
0.8	0,00	0,22	0,00	0,14	0,14	0,74	0,06	0,06	0,170
0.9	0,00	0,48	0,00	0,17	0,17	1,00	0,06	0,06	0,243
1.0	0,19	0,48	0,00	0,17	0,17	1,00	0,06	0,06	0,266

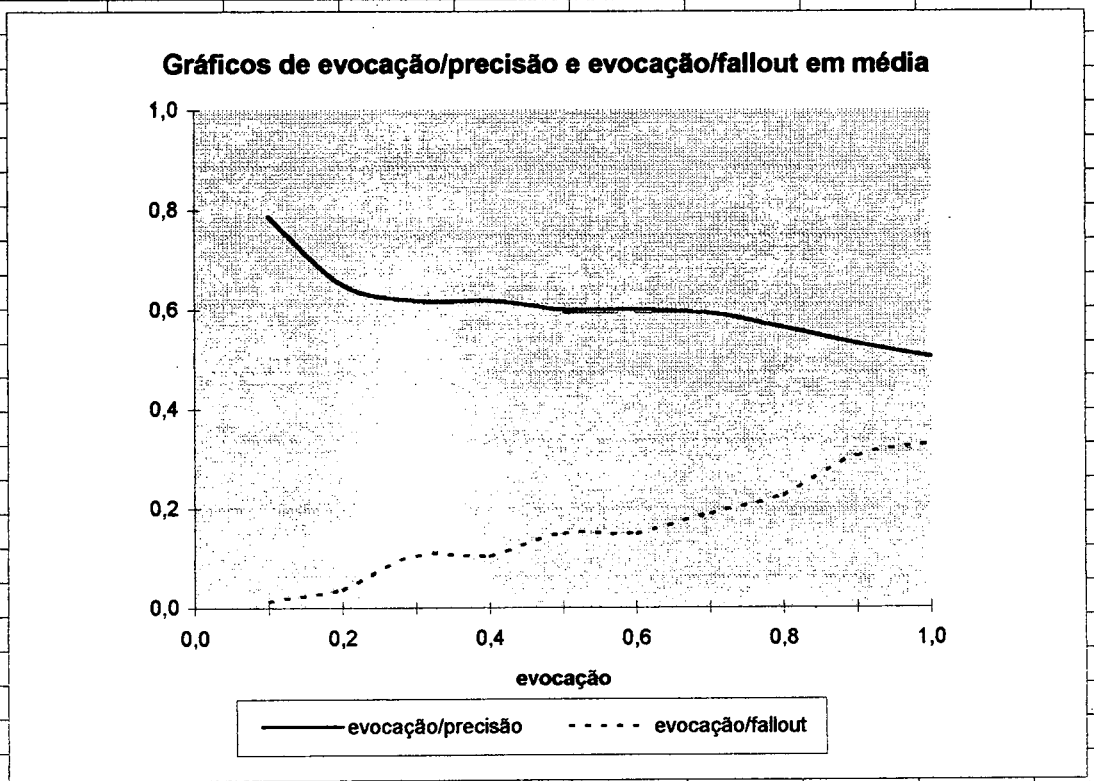
Gráficos de evocação/precisão e evocação/fallout em média



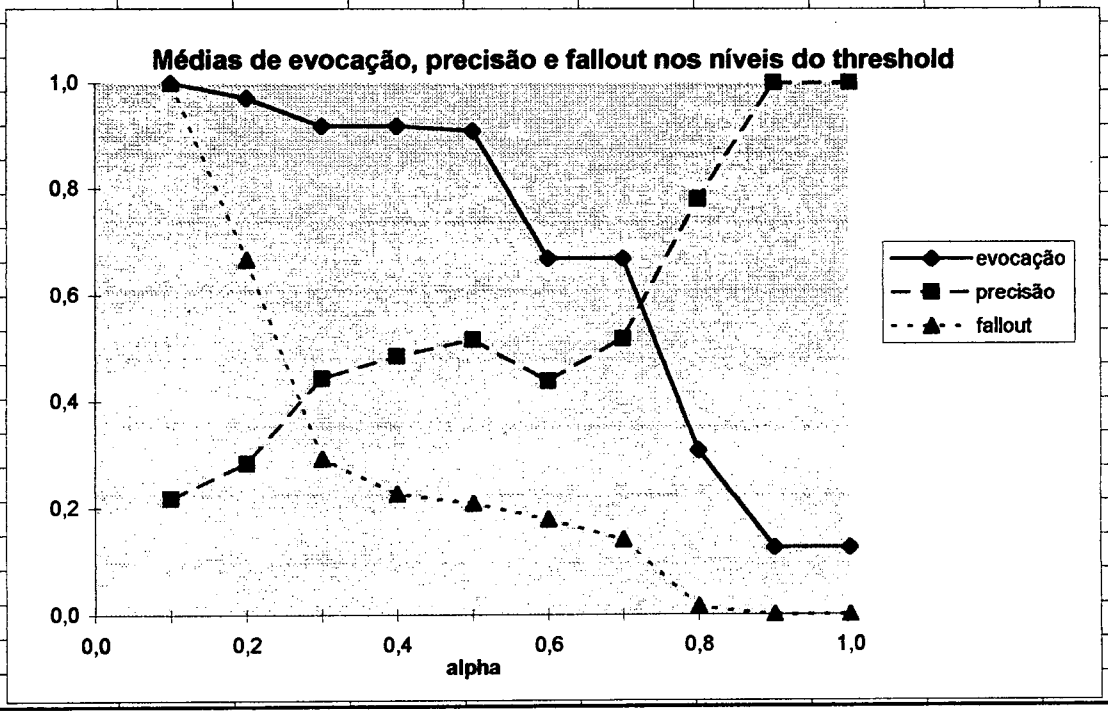
Radix			consultas não-controladas						
dicionário vazio			descritores controlados						
		Médias nos níveis do threshold (alpha)							
evocação									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	0,85	1,00	1,00	0,981
0.3	1,00	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,919
0.4	1,00	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,919
0.5	0,93	0,89	1,00	0,00	1,00	0,46	1,00	1,00	0,785
0.6	0,00	0,89	1,00	0,00	1,00	0,46	1,00	1,00	0,669
0.7	0,00	0,00	0,00	0,00	0,00	0,00	1,00	1,00	0,250
0.8	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
precisao									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,42	0,25	0,11	0,19	0,19	0,36	0,11	0,11	0,218
0.2	0,56	0,32	0,15	0,26	0,26	0,39	0,14	0,14	0,278
0.3	0,65	0,50	0,25	0,44	0,44	0,55	0,36	0,36	0,444
0.4	0,75	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,485
0.5	1,00	0,57	0,29	-	0,50	0,55	0,36	0,36	0,519
0.6	-	0,57	0,29	-	0,50	0,55	0,36	0,36	0,438
0.7	-	-	-	-	-	-	0,50	0,50	0,500
0.8	-	-	-	-	-	-	-	-	-
0.9	-	-	-	-	-	-	-	-	-
1.0	-	-	-	-	-	-	-	-	-
fallout									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	0,57	0,70	0,72	0,69	0,69	0,74	0,75	0,75	0,701
0.3	0,38	0,30	0,38	0,31	0,31	0,22	0,22	0,22	0,293
0.4	0,14	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,226
0.5	0,00	0,22	0,31	0,00	0,24	0,22	0,22	0,22	0,179
0.6	0,00	0,22	0,31	0,00	0,24	0,22	0,22	0,22	0,179
0.7	0,00	0,00	0,00	0,00	0,00	0,00	0,13	0,13	0,033
0.8	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000



Radix			consultas não-controladas						
dicionário vazio			descritores controlados						
	Médias da precisão nos níveis da evocação								
evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	0,50	1,00	1,00	0,67	0,57	0,57	0,789
0.2	1,00	0,64	0,50	0,67	0,67	0,60	0,57	0,57	0,653
0.3	1,00	0,64	0,50	0,54	0,54	0,60	0,57	0,57	0,620
0.4	1,00	0,64	0,50	0,54	0,54	0,60	0,57	0,57	0,620
0.5	1,00	0,64	0,50	0,54	0,54	0,45	0,57	0,57	0,601
0.6	1,00	0,64	0,50	0,54	0,54	0,45	0,57	0,57	0,601
0.7	1,00	0,64	0,50	0,54	0,54	0,39	0,57	0,57	0,594
0.8	1,00	0,62	0,29	0,54	0,54	0,39	0,57	0,57	0,565
0.9	1,00	0,39	0,29	0,54	0,54	0,36	0,57	0,57	0,533
1.0	0,79	0,39	0,29	0,54	0,54	0,36	0,57	0,57	0,506
	Médias do fallout nos níveis da evocação								
evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,00	0,00	0,00	0,00	0,00	0,04	0,03	0,03	0,013
0.2	0,00	0,07	0,00	0,03	0,03	0,09	0,03	0,03	0,035
0.3	0,00	0,11	0,09	0,14	0,14	0,17	0,09	0,09	0,104
0.4	0,00	0,11	0,09	0,14	0,14	0,17	0,09	0,09	0,104
0.5	0,00	0,15	0,09	0,17	0,17	0,43	0,09	0,09	0,149
0.6	0,00	0,15	0,09	0,17	0,17	0,43	0,09	0,09	0,149
0.7	0,00	0,15	0,09	0,17	0,17	0,74	0,09	0,09	0,188
0.8	0,00	0,15	0,31	0,21	0,21	0,74	0,09	0,09	0,225
0.9	0,00	0,52	0,31	0,21	0,21	1,00	0,09	0,09	0,304
1.0	0,19	0,52	0,31	0,21	0,21	1,00	0,09	0,09	0,328

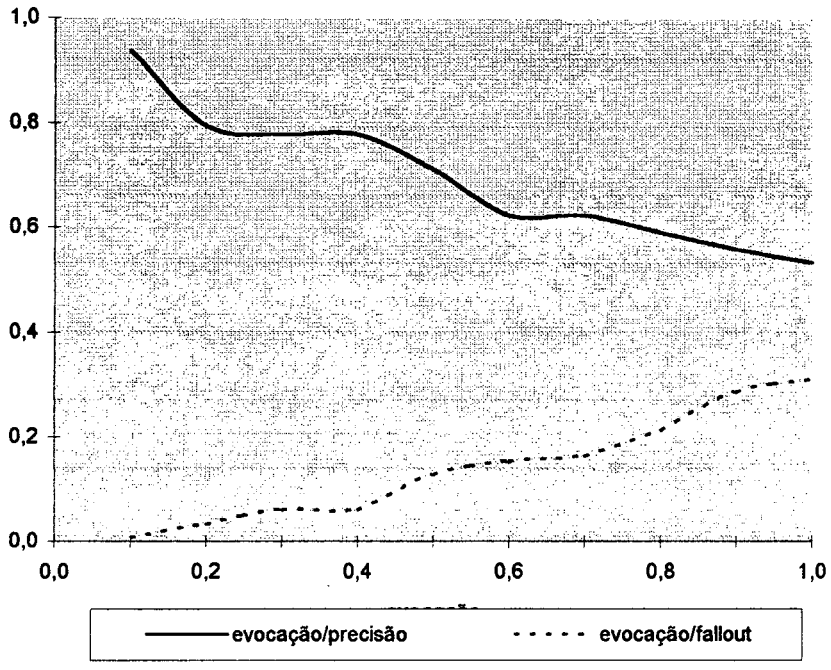


Radix			consultas não-controladas						
dicionário não-vazio			descritores controlados						
		Médias nos níveis do threshold (alpha)							
evocação									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	0,77	1,00	1,00	0,971
0.3	1,00	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,919
0.4	1,00	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,919
0.5	0,93	0,89	1,00	1,00	1,00	0,46	1,00	1,00	0,910
0.6	0,00	0,89	1,00	0,00	1,00	0,46	1,00	1,00	0,669
0.7	0,00	0,89	1,00	0,00	1,00	0,46	1,00	1,00	0,669
0.8	0,00	0,00	0,00	0,00	0,00	0,46	1,00	1,00	0,308
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,50	0,125
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,50	0,125
precisao									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,42	0,25	0,11	0,19	0,19	0,36	0,11	0,11	0,218
0.2	0,56	0,32	0,15	0,26	0,26	0,40	0,16	0,16	0,284
0.3	0,65	0,50	0,25	0,44	0,44	0,55	0,36	0,36	0,444
0.4	0,75	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,485
0.5	1,00	0,57	0,29	0,50	0,50	0,55	0,36	0,36	0,516
0.6	-	0,57	0,29	-	0,50	0,55	0,36	0,36	0,438
0.7	-	0,57	0,29	-	0,50	0,75	0,50	0,50	0,518
0.8	-	-	-	-	-	1,00	0,67	0,67	0,780
0.9	-	-	-	-	-	-	1,00	1,00	1,000
1.0	-	-	-	-	-	-	1,00	1,00	1,000
fallout									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	0,57	0,70	0,72	0,69	0,69	0,65	0,66	0,66	0,668
0.3	0,38	0,30	0,38	0,31	0,31	0,22	0,22	0,22	0,293
0.4	0,14	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,226
0.5	0,00	0,22	0,31	0,24	0,24	0,22	0,22	0,22	0,209
0.6	0,00	0,22	0,31	0,00	0,24	0,22	0,22	0,22	0,179
0.7	0,00	0,22	0,31	0,00	0,24	0,09	0,13	0,13	0,140
0.8	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,06	0,015
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000

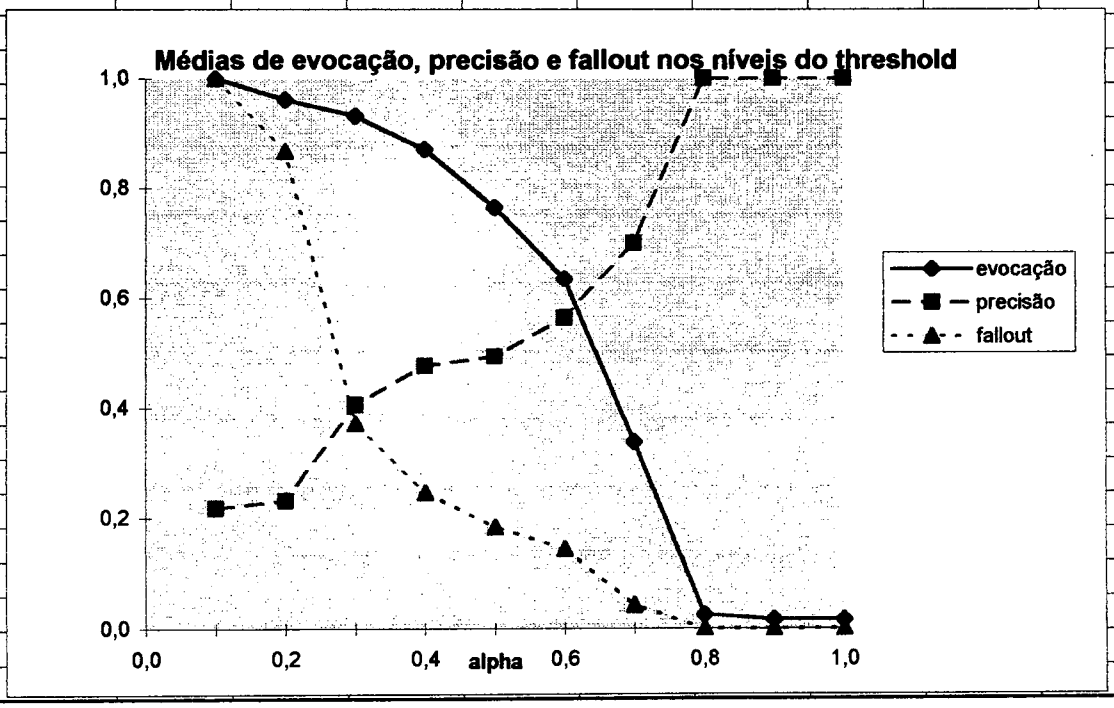


Radix			consultas não-controladas						
dicionário não-vazio			descritores controlados						
	Médias da precisão nos níveis da evocação								
evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	0,50	1,00	1,00	1,00	1,00	1,00	0,938
0.2	1,00	0,64	0,50	0,67	0,55	1,00	1,00	1,00	0,795
0.3	1,00	0,64	0,50	0,54	0,55	1,00	1,00	1,00	0,779
0.4	1,00	0,64	0,50	0,54	0,55	1,00	1,00	1,00	0,779
0.5	1,00	0,64	0,50	0,54	0,55	0,47	1,00	1,00	0,713
0.6	1,00	0,64	0,50	0,54	0,55	0,42	0,67	0,67	0,624
0.7	1,00	0,64	0,50	0,54	0,55	0,42	0,67	0,67	0,624
0.8	1,00	0,62	0,29	0,54	0,55	0,39	0,67	0,67	0,591
0.9	1,00	0,41	0,29	0,54	0,54	0,36	0,67	0,67	0,560
1.0	0,79	0,41	0,29	0,54	0,54	0,36	0,67	0,67	0,534
	Médias do fallout nos níveis da evocação								
evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,00	0,00	0,06	0,00	0,00	0,00	0,00	0,00	0,008
0.2	0,00	0,07	0,06	0,03	0,10	0,00	0,00	0,00	0,033
0.3	0,00	0,11	0,09	0,14	0,14	0,00	0,00	0,00	0,060
0.4	0,00	0,11	0,09	0,14	0,14	0,00	0,00	0,00	0,060
0.5	0,00	0,15	0,09	0,17	0,14	0,35	0,06	0,06	0,128
0.6	0,00	0,15	0,09	0,17	0,17	0,52	0,06	0,06	0,153
0.7	0,00	0,15	0,09	0,17	0,17	0,61	0,06	0,06	0,164
0.8	0,00	0,15	0,31	0,21	0,17	0,74	0,06	0,06	0,213
0.9	0,00	0,48	0,31	0,21	0,17	1,00	0,06	0,06	0,286
1.0	0,19	0,48	0,31	0,21	0,17	1,00	0,06	0,06	0,310

Gráficos de evocação/precisão e evocação/fallout em média

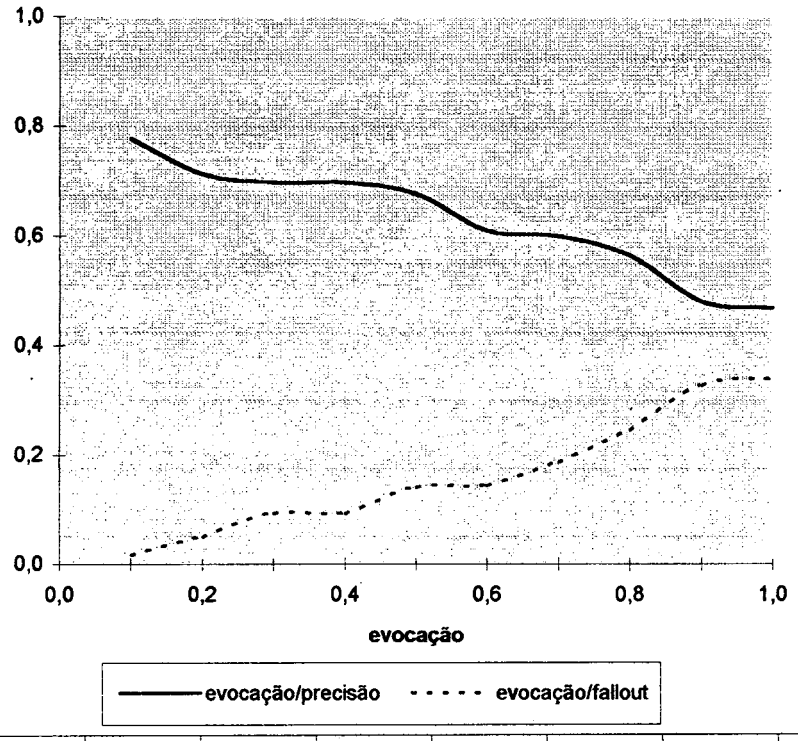


Radix			consultas não-controladas						
dicionário vazio			descritores não-controlados						
		Médias nos níveis do threshold (alpha)							
evocação									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	0,69	1,00	1,00	0,961
0.3	1,00	1,00	1,00	1,00	1,00	0,46	1,00	1,00	0,933
0.4	0,87	0,89	0,75	1,00	1,00	0,46	1,00	1,00	0,871
0.5	0,87	0,89	0,75	0,29	0,86	0,46	1,00	1,00	0,765
0.6	0,27	0,89	0,75	0,00	0,71	0,46	1,00	1,00	0,635
0.7	0,20	0,00	0,50	0,00	0,00	0,00	1,00	1,00	0,338
0.8	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,025
0.9	0,13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,016
1.0	0,13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,016
precisao									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,42	0,25	0,11	0,19	0,19	0,36	0,11	0,11	0,218
0.2	0,43	0,26	0,11	0,20	0,20	0,35	0,15	0,15	0,231
0.3	0,63	0,38	0,19	0,33	0,44	0,55	0,36	0,36	0,405
0.4	0,87	0,53	0,20	0,47	0,47	0,55	0,36	0,36	0,476
0.5	1,00	0,62	0,23	0,33	0,50	0,55	0,36	0,36	0,494
0.6	1,00	0,89	0,23	-	0,56	0,55	0,36	0,36	0,564
0.7	1,00	-	1,00	-	-	-	0,40	0,40	0,700
0.8	1,00	-	-	-	-	-	-	-	1,000
0.9	1,00	-	-	-	-	-	-	-	1,000
1.0	1,00	-	-	-	-	-	-	-	1,000
fallout									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	0,95	0,96	0,97	0,97	0,97	0,74	0,69	0,69	0,868
0.3	0,43	0,56	0,53	0,48	0,31	0,22	0,22	0,22	0,371
0.4	0,10	0,26	0,38	0,28	0,28	0,22	0,22	0,22	0,245
0.5	0,00	0,19	0,31	0,14	0,17	0,22	0,22	0,22	0,184
0.6	0,00	0,04	0,31	0,00	0,14	0,22	0,22	0,22	0,144
0.7	0,00	0,00	0,00	0,00	0,00	0,00	0,17	0,17	0,043
0.8	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000

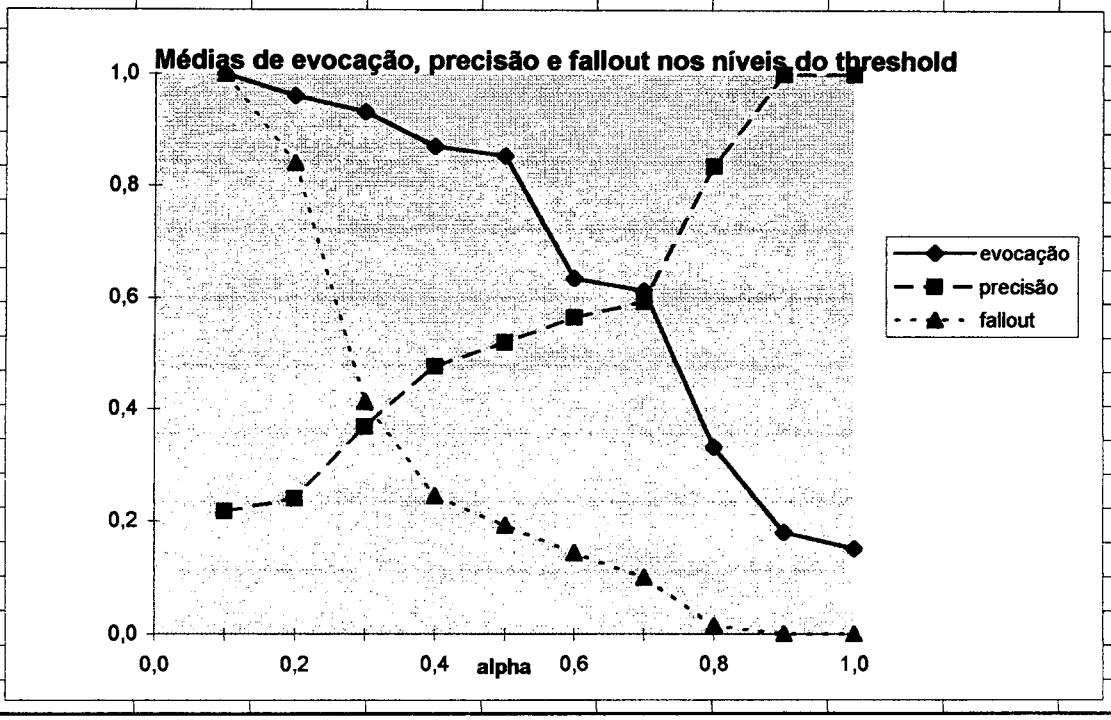


Radix			consultas não-controladas						
dicionário vazio			descritores não-controlados						
	Médias da precisão nos níveis da evocação								
evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	0,55	1,00	0,67	0,50	0,50	0,778
0.2	1,00	0,89	1,00	0,55	0,67	0,60	0,50	0,50	0,714
0.3	1,00	0,89	1,00	0,55	0,67	0,60	0,44	0,44	0,699
0.4	1,00	0,89	1,00	0,55	0,67	0,60	0,44	0,44	0,699
0.5	1,00	0,89	1,00	0,55	0,67	0,44	0,44	0,44	0,679
0.6	1,00	0,89	0,50	0,55	0,63	0,44	0,44	0,44	0,611
0.7	1,00	0,89	0,50	0,55	0,63	0,36	0,44	0,44	0,601
0.8	1,00	0,89	0,25	0,55	0,60	0,36	0,44	0,44	0,566
0.9	0,88	0,47	0,25	0,54	0,47	0,36	0,44	0,44	0,481
1.0	0,79	0,47	0,25	0,54	0,47	0,36	0,44	0,44	0,470
	Médias do fallout nos níveis da evocação								
evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,00	0,00	0,00	0,03	0,00	0,04	0,03	0,03	0,016
0.2	0,00	0,04	0,00	0,14	0,07	0,09	0,03	0,03	0,050
0.3	0,00	0,04	0,00	0,14	0,07	0,17	0,16	0,16	0,093
0.4	0,00	0,04	0,00	0,14	0,07	0,17	0,16	0,16	0,093
0.5	0,00	0,04	0,09	0,17	0,07	0,43	0,16	0,16	0,140
0.6	0,00	0,04	0,09	0,17	0,10	0,43	0,16	0,16	0,144
0.7	0,00	0,04	0,09	0,17	0,10	0,78	0,16	0,16	0,188
0.8	0,00	0,04	0,38	0,17	0,14	0,91	0,16	0,16	0,245
0.9	0,10	0,37	0,38	0,17	0,28	1,00	0,16	0,16	0,328
1.0	0,19	0,37	0,38	0,17	0,28	1,00	0,16	0,16	0,339

Gráficos de evocação/precisão e evocação/fallout em média



Radix			consultas não-controladas						
dicionário não-vazio			descritores não-controlados						
		Médias nos níveis do threshold (alpha)							
evocação									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	1,00	1,00	1,00	1,00	0,69	1,00	1,00	0,961
0.3	1,00	1,00	1,00	1,00	1,00	0,46	1,00	1,00	0,933
0.4	0,87	0,89	0,75	1,00	1,00	0,46	1,00	1,00	0,871
0.5	0,87	0,89	0,75	1,00	0,86	0,46	1,00	1,00	0,854
0.6	0,27	0,89	0,75	0,00	0,71	0,46	1,00	1,00	0,635
0.7	0,20	0,78	0,75	0,00	0,71	0,46	1,00	1,00	0,613
0.8	0,20	0,00	0,00	0,00	0,00	0,46	1,00	1,00	0,333
0.9	0,13	0,00	0,00	0,00	0,00	0,31	0,50	0,50	0,180
1.0	0,13	0,00	0,00	0,00	0,00	0,08	0,50	0,50	0,151
precisao									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,42	0,25	0,11	0,19	0,19	0,36	0,11	0,11	0,218
0.2	0,43	0,26	0,11	0,20	0,20	0,38	0,17	0,17	0,240
0.3	0,63	0,38	0,19	0,33	0,41	0,43	0,29	0,29	0,369
0.4	0,87	0,53	0,20	0,47	0,47	0,55	0,36	0,36	0,476
0.5	1,00	0,62	0,23	0,54	0,50	0,55	0,36	0,36	0,520
0.6	1,00	0,89	0,23	-	0,56	0,55	0,36	0,36	0,564
0.7	1,00	0,88	1,00	-	0,63	0,55	0,36	0,36	0,593
0.8	1,00	-	-	-	-	1,00	0,67	0,67	0,835
0.9	1,00	-	-	-	-	1,00	1,00	1,00	1,000
1.0	1,00	-	-	-	-	1,00	1,00	1,00	1,000
faliout									
alpha	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	0,95	0,96	0,97	0,97	0,97	0,65	0,63	0,63	0,841
0.3	0,43	0,56	0,53	0,48	0,34	0,35	0,31	0,31	0,414
0.4	0,10	0,26	0,38	0,28	0,28	0,22	0,22	0,22	0,245
0.5	0,00	0,19	0,31	0,21	0,17	0,22	0,22	0,22	0,193
0.6	0,00	0,04	0,31	0,00	0,14	0,22	0,22	0,22	0,144
0.7	0,00	0,04	0,00	0,00	0,10	0,22	0,22	0,22	0,100
0.8	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,06	0,015
0.9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
1.0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000



Radix			consultas não-controladas						
dicionário não-vazio			descritores não-controlados						
	Médias da precisão nos níveis da evocação								
evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
0.2	1,00	0,89	1,00	0,60	0,67	1,00	1,00	1,00	0,895
0.3	1,00	0,89	1,00	0,60	0,67	1,00	1,00	1,00	0,895
0.4	1,00	0,89	1,00	0,60	0,67	1,00	1,00	1,00	0,895
0.5	1,00	0,89	1,00	0,60	0,67	0,41	1,00	1,00	0,821
0.6	1,00	0,89	1,00	0,60	0,63	0,41	0,67	0,67	0,734
0.7	1,00	0,89	1,00	0,60	0,63	0,38	0,67	0,67	0,730
0.8	1,00	0,89	0,25	0,60	0,60	0,37	0,67	0,67	0,631
0.9	0,88	0,47	0,25	0,54	0,47	0,36	0,67	0,67	0,539
1.0	0,79	0,47	0,25	0,54	0,47	0,36	0,67	0,67	0,528
	Médias do fallout nos níveis da evocação								
evocação	C1	C2	C3	C4	C5	C6	C7	C8	média
0.1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,000
0.2	0,00	0,04	0,00	0,10	0,07	0,00	0,00	0,00	0,026
0.3	0,00	0,04	0,00	0,10	0,07	0,00	0,00	0,00	0,026
0.4	0,00	0,04	0,00	0,10	0,07	0,00	0,00	0,00	0,026
0.5	0,00	0,04	0,00	0,10	0,07	0,57	0,06	0,06	0,113
0.6	0,00	0,04	0,00	0,14	0,10	0,57	0,06	0,06	0,121
0.7	0,00	0,04	0,00	0,14	0,10	0,70	0,06	0,06	0,138
0.8	0,00	0,04	0,38	0,14	0,14	0,83	0,06	0,06	0,206
0.9	0,10	0,37	0,38	0,17	0,28	1,00	0,06	0,06	0,303
1.0	0,19	0,37	0,38	0,17	0,28	1,00	0,06	0,06	0,314

Gráficos de evocação/precisão e evocação/fallout em média