ANGELO ALFREDO SUCOLOTTI

RECUPERAÇÃO DE INFORMAÇÃO EM BASES TEXTUAIS: UMA ABORDAGEM BASEADA EM LÓGICA PARACONSISTENTE

FLORIANÓPOLIS – SC

2001

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM
CIÊNCIA DA COMPUTAÇÃO

Angelo Alfredo Sucolotti

RECUPERAÇÃO DE INFORMAÇÃO EM BASES TEXTUAIS: UMA ABORDAGEM BASEADA EM LÓGICA PARACONSISTENTE

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Orientador:
Prof. Dr. Murilo Silva de Camargo

Florianópolis, novembro/2001.

RECUPERAÇÃO DE INFORMAÇÃO EM BASES TEXTUAIS: UMA ABORDAGEM BASEADA EM LÓGICA PARACONSISTENTE

Angelo Alfredo Sucolotti

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação (Sistemas Computacionais) e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

Prof. Dr. Eng. Fernando A. Ostuni Gauthier
(Coordenador do Curso)

Banca Examinadora:

Prof. Dr. Eng. Murilo Silva de Camargo
(Orientador e Presidente da Banca)

Prof. Dr. Roberto Willrich

Prof. Dr. Vitório Bruno Mazzola

AGRADECIMENTOS

A todos que de alguma forma estiveram envolvidos com essa pesquisa.

Ao meu orientador, Prof. Dr. Murilo Silva de Camargo, pelo exemplo de pessoa e de profissional que me transmitiu durante esta caminhada. Obrigado por estimular meu pensar e por acreditar em minha capacidade.

A todos os amigos que certamente irão se reconhecer neste agradecimento, pois sempre contei com solidariedade, envolvimento e auxílio.

Aos familiares, que souberam compreender minhas angústias e ausências.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

No presente estudo é apresentada uma abordagem para recuperação de informação em bases textuais usando um modelo baseado em lógica paraconsistente. A idéia base utilizada na abordagem proposta é a expansão do conjunto de termos indexados na *query* com termos relacionados, obtidos de um *thesaurus*, utilizando como base o algoritmo "para-analisador".

A utilização da lógica paraconsistente é justificada pela facilidade proporcionada por esta, no tratamento de situações que envolvem incertezas, paradoxos, inconsistências e vagacidade (*vagueness*), uma vez que bases textuais e *queries* em sua maioria são repletas destas situações.

Esse trabalho revelou que é possível e provável que a implementação de um modelo com essas características proporcione melhora da qualidade dos termos que constituem o *thesaurus*.

Os resultado desta se constituem assim, numa contribuição científica original para a área, uma vez que não existe nenhum estudo específico e detalhado para avaliação de técnicas de lógica paraconsistente para recuperação de informação.

ABSTRACT

This work introducing an approach to retrieval of information in textual bases using a model based in paraconsistent logic. The base idea utilized in the approach offered is the expansion of the conjoined of indexed terms in the query with related, obtained of a thesaurus, utilizing as base the algorithm "para-analyzer".

The utilization of the paraconsistent logic is justificated for the facility offered for it in the treatment of situation that involve uncertainty, paradox, inconsistency and vagueness, since textual bases and queries in majority are replete these situations.

This research exposed that is possible and provable the implementation of a model with this characteristic provides improvement of the quality of the terms that constitute the thesaurus.

The results this is constituted in an original scientific contribution for an area, since there isn't any specific study and detailed to valuation of technique of paraconsistent logic to recovery of information.

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

No atual contexto mundial, evidencia-se uma imensa quantidade de informações e a disseminação dessas com grande rapidez nas mais diversas áreas do conhecimento.

A utilização dos recursos da informática tem grande parcela de responsabilidade pelas grandiosas conquistas, principalmente no que se refere à difusão de informações.

Percebe-se no que tange ao processo de recuperação de informação que há lacunas. A pesquisa que segue trata-se de um estudo sobre a recuperação de informação em bases textuais utilizando os recursos da lógica paraconsistente.

A realização deste se deve as dificuldades encontradas na busca por métodos de recuperação de informação eficientes que tratem com situações que apresentam contradições.

A principal motivação para apresentação da proposta desta pesquisa se deve ao crescimento considerável da utilização de base textuais, e isso acarreta a necessidade de fornecer mecanismos eficientes para recuperação de informação. Um dos fatores que contribuíram para esse crescimento pode ser atribuído à difusão da Word Wide Web (WWW), bem como a crescente utilização de bibliotecas digitais, advindas do aperfeiçoamento tecnológico.

Em bases textuais, o crescimento dos objetos armazenados e o grande volume de informação exigem processos de recuperação cada vez mais sofisticados. Em função disso, a recuperação de informação apresenta a cada dia novos desafios e torna-se cada vez mais importante.

A área de recuperação de informação abrange o estudo de métodos para fornecer aos usuários o pequeno subconjunto de informação relevante às suas necessidades. A princípio informação pode aceitar muitas formas como imagens, sons, texto, entre outros. Neste estudo será utilizado sistema de informação baseado em texto.

A razão pela qual será utilizada lógica paraconsistente como uma abordagem para recuperação de informação em bases textuais decorre da facilidade proporcionada por esta no tratamento de incertezas, paradoxos, inconsistências e vagacidade (*vagueness)*.

Propõem-se nesta pesquisa o estudo da viabilidade de um modelo com essas características, que deverá proporcionar um raciocínio mais flexível, capaz de tratar com incertezas, incompletude e falta de informações, o que certamente acarretará um aumento considerável na performance de sistemas de recuperação de informação.

1.2 METODOLOGIA

Para a realização deste trabalho, foram pesquisadas diversas bibliografias, tais como: livros, dissertações, teses, artigos, relatórios técnicos, documentos oficiais de congressos, Workshops e sites da Internet. O material utilizado foi obtido principalmente por meio de pesquisa em bibliotecas digitais e Comut.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado da seguinte forma: O capítulo 2 apresenta uma revisão sobre os principais aspectos relacionados à recuperação de informação, tais como motivação, apresentação do problema, recuperação de informação versus recuperação de dados, definição formal de um modelo, modelos clássicos, modelos alternativos, métricas para avaliação, entre outros.

Os capítulos 3 e 4 apresentam uma revisão bibliográfica sobre os abrangentes temas de lógica paraconsistente e *thesaurus*, respectivamente.

No capítulo 5 realiza-se a discussão dos resultados onde é apresentado um modelo para recuperação de informação em bases textuais: uma abordagem paraconsistente que proporcionará um raciocínio mais flexível capaz de lidar com incertezas, incompletude e falta de informações, o que certamente proporcionará um aumento considerável da performance de sistemas de recuperação de informação.

Finalmente, apresentam-se as conclusões obtidas e as referências bibliográficas utilizadas nesta pesquisa.

2 RECUPERAÇÃO DE INFORMAÇÃO

2.1 INTRODUÇÃO

Neste capítulo apresentar-se-á a definição do problema no que se refere à recuperação de informação. Também serão demonstradas as principais abordagens utilizadas no processo de recuperação de informação.

2.2 MOTIVAÇÃO

Inicialmente recuperação de informação era uma área dita de interesse restrito a algumas áreas, principalmente a bibliotecas e sistemas especialistas. Entretanto com o crescimento cada vez maior de conhecimento disponível em meios digitais, tais como enciclopédias e principalmente com a difusão da Word Wide Web, este conhecimento excede em muito a capacidade humana em recuperar e analisar estas informações, o que torna cada vez mais necessário proporcionar mecanismos precisos e eficazes para recuperação de informação.

Atualmente pesquisas na área de recuperação de informação abrangem inúmeras novas áreas, tais como: modelagem, interfaces com o usuário, arquitetura de sistemas, visualização dos dados, filtragem, linguagens, etc.

2.3 APRESENTAÇÃO DO PROBLEMA

À primeira vista, armazenamento e recuperação de informação são tarefas bastante simples. Considerando há existência de uma coleção de documentos, na qual uma pessoa necessite buscar informações, sendo que nesta coleção encontram-se documentos relevantes (úteis) e não relevantes a sua necessidade. Uma solução para recuperar esses documentos ditos relevantes a necessidade dessa pessoa seria a leitura de todos os documentos dessa coleção. Como resultado desta operação obter-se-ia uma recuperação "perfeita", contudo conclui-se que essa solução é impraticável. Desse modo percebe-se que o problema de recuperação de informação é algo um tanto quanto complexo de ser tratado, evidenciando a necessidade da realização de pesquisas para solucionar esta dificuldade.

O propósito de estratégia de recuperação automática de informação está em recuperar todos os documentos relevantes a necessidade do usuário.

Para melhor elucidar o problema de recuperação de informação é necessário à compreensão da existência de diferenças entre recuperação de dados e recuperação de informação, as quais serão apresentadas a seguir.

2.4 RECUPERAÇÃO DE DADOS VERSUS RECUPERAÇÃO DE INFORMAÇÃO

Recuperação de dados em um sistema de recuperação de informação (IR), consiste principalmente em determinar quais documentos de uma coleção contém a palavra-chave da *query* realizada pelo usuário, ressalta-se que um dos métodos mais utilizados para recuperação de informação, em bases textuais, são pesquisas baseadas em palavras-chaves. De maneira geral, isto não é suficiente para satisfazer a necessidade do usuário. Na verdade o usuário de um sistema de recuperação de informação necessita efetivamente recuperar informação sobre um assunto específico e não apenas recuperar dados que satisfaçam a uma dada *query*.

Query é a expressão de necessidade de informação do usuário de um sistema de recuperação de informação. Neste estudo considera-se uma *query* um único evento; isto é, se dois usuários submetem a mesma *query* ou se a mesma *query* é submetida pelo mesmo usuário em duas ocasiões diferentes, as duas *query* são consideradas diferentes. Uma *query* é submetida ao sistema que objetiva encontrar informações relevantes para a necessidade de informação expressada nesta *query*.

Considera-se relevância como o julgamento subjetivo do usuário relacionado ao documento relatado para uma única expressão de necessidade de informação. Relevância relacionada entre uma *query* e um documento conta com a satisfação da necessidade de informação do usuário. Tal satisfação é subjetiva, diferentes usuários podem dar diferentes julgamentos de relevância para uma dada relação *query* e documento, o que também pode variar, pois o mesmo usuário em ocasiões diferentes pode atribuir valores de relevância diferentes para um mesmo documento.

Uma linguagem de recuperação de dados tem como objetivo recuperar todos documentos que satisfaçam perfeitamente as condições definidas, tal como nas expressões regulares ou na álgebra relacional, o que é confirmado por [Ribeiro-Neto e Baeza-Yates, 1999]. Em função disso um único termo errado entre vários outros

corretos podem causar falhas em um sistema de recuperação de dados.

Já para um sistema de recuperação de informação, a recuperação pode ser inexata e é possível que pequenos erros não sejam percebidos. A principal razão para essa diferença é que recuperação de informação geralmente trata com dados (textos) em linguagem natural que nem sempre são bem estruturados e podem ser semanticamente ambíguos. Por outro lado, um sistema de recuperação de dados como, por exemplo, uma base de dados relacional, trata seus dados com estrutura e semântica bem definidas.

Recuperação de dados continua proporcionando uma solução para os usuários de um sistema de banco de dados, mas não soluciona problemas de recuperação de informação inseridas em um determinado contexto. O que pode ser exemplificando no caso de uma consulta em que o usuário necessite recuperar informações sobre veículos acidentados, utilizando um sistema de recuperação de dados, no qual não seriam recuperadas informações relacionadas a automóveis batidos.

Para atender de forma efetiva as necessidades de informação de um usuário, um sistema de recuperação de informação deve de alguma maneira interpretar o conteúdo de dados, itens de informação, em uma coleção de documentos e assim classificar conforme um grau de relevância para a *query* do usuário.

Esta interpretação de um documento implica em satisfazer extraindo sintática e semanticamente as informações de documentos e utilizá-los para condizer com a necessidade de informação do usuário. A dificuldade não está apenas no conhecimento para extração dessas informações, mas também no conhecimento para utilização na decisão de relevância.

Estabelecer graus de relevância de um documento para uma *query* é intelectualmente possível para os seres humanos. Para que um computador realize esta tarefa precisa-se construir um modelo em que decisões de relevância possam ser estabelecidas.

De fato, a primeira barreira de um sistema de recuperação de informação está em recuperar todos os documentos que são relevantes para a *query* do usuário, bem como recuperar alguns documentos não totalmente relevantes, ou seja, que satisfaçam parcialmente a necessidade do usuário.

2.5 DEFINIÇÃO FORMAL DE UM MODELO PARA RECUPERAÇÃO DE INFORMAÇÃO

Um modelo para recuperação de informação é um conjunto de premissas e um algoritmo para classificação de documentos, com o objetivo de atender a uma *query* de um usuário. Uma definição formal é expressa em [Ribeiro-Neto e Baeza-Yates, 1999] aonde um modelo de recuperação de informação (IR) é uma quádrupla [**D**, **Q**, F, R(q_i, d_j)] onde:

D é um conjunto composto de visões lógicas dos documentos de uma coleção;

Q é um conjunto composto de visões lógicas para as necessidades de informações do usuário;

F é uma estrutura para modelar documentos e *query*, e seus relacionamentos;

R(q_i, d_j) é uma função de classificação, a qual associa-se um número real que uma *query* $q_i \in$ **Q** e uma representação do documento $d_j \in$ **D.** Tal classificação define uma ordem entre documentos que atendem a *query* q_i.

Várias abordagens são utilizadas no contexto de recuperação de informação, havendo uma classificação dessas abordagens de acordo com suas características. A seguir será apresentada uma breve explanação sobre estas abordagens.

2.6 MODELOS CLÁSSICOS PARA RECUPERAÇÃO DE INFORMAÇÃO

Existem três modelos de recuperação de informação, que são classificados como modelos clássicos, são eles: booleano, vetor e probabilístico.

Para facilitar o entendimento destes modelos, alguns conceitos básicos serão apresentados.

2.6.1 Conceitos Básicos

2.6.1.1 Termos indexados

Os modelos clássicos em recuperação de informação consideram que cada documento é descrito por um conjunto de palavras chaves chamadas termos indexados. Um termo indexado nada mais é do que uma palavra que consegue representar a semântica de um ou vários documentos contidos em uma coleção de documentos, assim termos indexados são utilizados para indexar ou resumir o conteúdo de um documento.

Para representar estes termos indexados, geralmente são utilizados substantivos, pois eles mesmos apresentam sentido próprio, já a utilização de advérbios, adjetivos ou conectivos são de menor proveito, pois estes desempenham um papel de complemento.

2.6.1.2 Relevância

O conceito de relevância é o conceito fundamental em recuperação de informação. Algumas tentativas de definição formal do que significa relevância em recuperação de informação são encontradas em [Seracevic, 1970], [Copper, 1971] e [Mizzaro, 1996].

Observa-se que nem todos os termos são relevantes para representar o conteúdo de um documento, alguns representam melhor a semântica de um documento do que outros.

Desta forma, evidencia-se que cada termo indexado possui um certo grau ou peso de relevância para descrever os documentos de uma coleção. A utilização de um peso atribuído ao termo indexado faz-se necessário para demonstrar a relevância deste termo em representar a semântica do documento.

Seja ki um termo indexado, dj um documento e wij ≥ 0 o peso associado ao par (ki, dj).

Definição Formal: Seja t o número de termos indexados pelo sistema, e k_i um termo indexado qualquer. $K = \{k_1,...,k_t\}$ é o conjunto de todos os termos indexados. $W_{ij} > 0$ é o peso associado com cada termo indexado k_i, do documento d_j. Os termos indexados que não aparecem no documento terão $w_{ij} = 0$. Com o documento d_j é então associado um vetor $\mathbf{d}_j$ representado por $\mathbf{d}_j = (w1,j, w2,j,...w_{t,j})$.

De modo que definir se um termo é relevante para expressar a semântica de um documento é uma tarefa que apresenta grande complexidade.

2.6.2 Modelo Booleano

O modelo booleano para recuperação de informação é baseado na teoria de conjuntos e álgebra booleana.

As principais vantagens da utilização deste modelo estão baseadas no seu formalismo claro e na sua simplicidade. Os primeiros sistemas comerciais para bibliotecas foram baseados neste modelo.

A recuperação de informação deste modelo é baseada em critérios binários, lógica clássica, isto é, os documentos são classificados em dois grupos: os relevantes e não relevantes a *query* do usuário. Nesta forma de recuperação, como resposta a uma

solicitação do usuário obter-se-á um conjunto resposta muito grande ou muito pequeno.

A modelagem das *queries* dos usuários em expressões booleanas é um trabalho complexo, pois estas expressões possuem semântica precisa, acrescido do fato de que muitos usuários não treinados em matemática apresentam dificuldade em traduzir suas necessidades de informação em expressões booleanas, essas dificuldades envolvem o trabalho com ordem de precedência dos operadores, uso de parênteses, entre outras.

Sistemas de recuperação de informação modernos não são baseados no modelo booleano, pois estes em função de suas características são comprovadamente pouco eficazes para recuperação de informação, o que é confirmado por [Ribeiro-Neto e Baeza-Yates, 1999].

Neste modelo não são utilizadas formas de classificação ou de atribuição de grau de relevância para os documentos recuperados pela *query* do usuário, isto é, não há nenhuma distinção entre os documentos recuperados.

[Ribeiro-Neto e Baeza-Yates, 1999] referem que a atribuição de um valor com função de identificar a similaridade entre os documentos contidos em uma coleção com os termos indexados contribuem significativamente na melhoria da performance de um sistema de recuperação de informação.

O modelo booleano é classificado como sendo o mais fraco dentre os modelos clássicos, em função de as suas características, principalmente por não permitir buscas parciais o que leva a uma performance pouco eficiente. Desta forma o modelo booleano esta mais relacionado com recuperação de dados do que com recuperação de informação.

2.6.3 Modelo Vetorial

É evidente que a utilização de pesos binários é bastante limitada, por esta razão o modelo vetorial propõe uma estrutura onde buscas parciais podem ser implementadas. Uma busca parcial pode ser caracterizada por uma solicitação de informação realizada pelo usuário que utiliza uma expressão composta que possua o conectivo AND. O que pode ser exemplificado na *query:* (Fórmula-1 and Senna), se algum documento satisfizer apenas parte da expressão, este pode ser recuperado pois satisfaz parcialmente a solicitação do usuário.

O modelo espaço vetorial ou simplesmente modelo vetorial trata documentos (***D***) e consultas (***Q***) como vetores em um espaço dimensional N, onde N é o número de

atributos indexados. O vetor de documentos é classificado conforme o seu co-seno de similaridade entre os vetores de termos **D** e **Q** ou através de alguma função de distância.

Este modelo utiliza atribuição de pesos não binários para os termos indexados da *query* e dos documentos. Esses termos indexados são utilizados para determinar o grau de similaridade entre a *query* e os documentos armazenados.

O modelo vetorial ao invés de separar os documentos em dois grupos relevantes e irrelevantes, como o modelo booleano, classifica os documentos de acordo com o seu grau de similaridade com os termos indexados.

Os pesos dos termos indexados podem ser calculados de várias maneiras diferentes. Em [Salton e McGill, 1983] são revistas várias dessas técnicas de atribuição de peso aos termos indexados.

Este modelo tem como princípios básicos suportar técnicas de *clustering*. *Clustering* é um grupo de documentos que satisfaçam à um conjunto de propriedades comuns.

O resultado de uma busca será armazenado em uma estrutura do tipo vetor, em ordem decrescente de similaridade dos documentos com a *query* realizada.

O conjunto resposta de uma busca utilizando esse método é mais preciso do que se for utilizado o método booleano, pois permite a recuperação de documentos que se aproximam das condições da *query*.

[Ribeiro-Neto e Baeza-Yates, 1999], referem que vários métodos de classificação tem sido propostos, mas parece ser consenso que a utilização deste modelo para recuperação de informação é uma boa alternativa. Confirmado por [Salton e McGill, 1983], quando referem que o modelo vetorial é um dos modelos mais utilizados em sistemas de recuperação de informação, isto é atribuída por sua simplicidade conceitual, sua aplicabilidade em bases genéricas e por sua eficiência.

2.6.4 Modelo Probabilístico

A primeira tentativa para desenvolver uma teoria probabilística para recuperação de informação foi realizada na década de 60 por [Maron e Kuhns, 1960] e [Miller, 1971] e desde então esta abordagem tem sido amplamente desenvolvida.

Uma característica comum de todos os modelos probabilísticos desenvolvidos para recuperação de informação é a utilização do princípio de classificação probabilística (PRP) [Robertson, 1977]. O princípio de classificação probabilística

afirma que a melhor performance de recuperação pode ser alcançada quando documentos são classificados de acordo com sua probabilidade de serem considerados relevantes para uma query.

Geralmente os modelos probabilísticos tem como seu espaço amostral o conjunto D X Q, onde Q representa o conjunto de todas consultas (*queries*) possíveis e D, o conjunto de todos os documentos de uma coleção. A diferença entre os vários modelos encontra-se na utilização de diferentes representações e descrições de consultas e documentos.

Este modelo tenta estimar a probabilidade de uma *query* realizada por um usuário encontrar um documento de uma coleção de documentos que seja relevante a sua *query*. O modelo assume que a probabilidade de um documento ser relevante depende somente da *query* e da descrição do documento.

O modelo considera que o conjunto de todos os documentos de uma coleção que satisfaçam a necessidade de informação do usuário forma um subconjunto resposta da *query* do usuário. Desta maneira os documentos contidos no conjunto resposta dito perfeito são relevantes a *query* do usuário. Aquela não contida neste conjunto resposta não é relevante.

As duas principais abordagens de sistemas de recuperação de informação probabilísticos são os modelos de relevância e inferência.

Modelos de relevância são baseados na evidência sobre quais documentos são relevantes para uma dada *query*. Estimar a probabilidade de relevância de cada documento de uma coleção é a maior dificuldade encontrada nesta abordagem, devido ao grande número de variáveis envolvidas na representação dos documentos em comparação a pequena quantidade de documentos relevantes.

Modelos de inferência utilizam-se de conceitos e técnicas advindas de outras áreas, tais como inteligência artificial e lógica fuzzy.

Do ponto de vista probabilístico, os modelos mais importantes são aqueles que tratam incertezas.

A vantagem dos modelos probabilísticos, segundo [Ribeiro-Neto e Baeza-Yates, 1999] em teoria, é que documentos são classificados em ordem decrescente de sua probabilidade de serem relevantes. Por outro lado, as desvantagens citadas são a necessidade de supor a separação inicial de documentos dentro de dois grupos os

relevantes e irrelevantes; o fato do método não adotar contador de freqüência com que cada termo indexado ocorre dentro do documento, isto é, todos os pesos são binários; e a adoção de hipótese de independência para termos indexados. Mas em situações reais não esta claro que o emprego desta seja inadequado.

De acordo com [Ribeiro-Neto e Baeza-Yates, 1999], alguns estudos teriam sugerido que o modelo probabilístico possui uma recuperação mais adequada em relação ao modelo vetorial. Porém estudos posteriores realizados por Salton e Buckley [apud Ribeiro-Neto e Baeza-Yates, 1999], descartam essa hipótese e justificam que o modelo vetorial é simples e rápido, e ainda tem a vantagem em termos de performance na recuperação de informação em coleções genéricas.

2.7 MODELOS ALTERNATIVOS PARA RECUPERAÇÃO DE INFORMAÇÃO

Além dos modelos clássicos, outras abordagens para recuperação de informação são propostas como sendo modelos alternativos e estes são classificados como: teóricos, algébricos e probabilísticos.

2.7.1 Modelos Teóricos

Os modelos alternativos teóricos são subdivididos em: Modelo Fuzzy e Modelo Booleano Estendido. Uma breve descrição destes modelos será apresentada.

2.7.1.1 Modelo fuzzy

Os modelos de conjuntos fuzzy para recuperação de informação propostos nos últimos anos utilizam conceitos básicos da teoria fuzzy ou lógica difusa.

Segundo [Klir e Yuan, 1995], a teoria fuzzy, introduzida por Loft Zadeh, trabalha com conjuntos difusos os quais não possuem limites precisos e a pertinência a estes conjuntos não se trata de um problema de afirmação ou negação, mas sim de um problema de grau. Esta lógica é baseada em uma lógica que trabalha com números dentro do intervalo [0,1], generalizando a lógica bi-valorada.

As características apresentadas por este novo paradigma proporcionam o tratamento de incerteza, vagacidade (*vagueness*) e incompletude nas informações presentes para resolução de um problema.

Observa-se que recuperar documentos através de *queries* que utilizem um conjunto de palavras chaves representa apenas parte da semântica real da *query* e dos

documentos, com isso tem-se uma busca parcial ou vaga, e como resultado obter-se-á uma recuperação aproximada ou vaga desta *query*.

Podemos modelar esse problema, considerando que cada termo da *query* define um conjunto fuzzy, e que cada documento tem um grau de pertinência nesse conjunto.

Recuperação de informação fuzzy é uma abordagem para modelar o processo de recuperação de informação que adota um dicionário para os termos indexados, fazendo com que o conjunto destes termos conseqüentemente seja expandido.

Conforme [Ribeiro-Neto e Baeza-Yates, 1999], atualmente esta abordagem para recuperação de informação tem sido pouco utilizada, esse modelo é pouco popular, sendo que a grande maioria dos experimentos nesta área tem utilizado pequenas coleções de dados.

2.7.1.2 Modelo booleano estendido

Vários modelos foram propostos utilizando os melhores recursos de outros modelos. O modelo booleano introduzido em 1983 por [Salton, Fox e Wu, 1983] é um destes modelos híbridos, pois é formado por características do modelo booleano clássico e do modelo vetorial.

O modelo booleano estendido ao contrário do modelo booleano clássico, permite buscas parciais e atribuição de pesos aos termos. Essa estratégia possibilita combinar formulários de *queries* booleanas com características do modelo vetorial.

Vários modelos foram propostos baseados na idéia de estender o modelo booleano clássico com os recursos do modelo vetorial, mesmo assim este modelo atualmente tem sido pouco utilizado, mas devido a sua estrutura bem definida pode vir a ser um modelo útil no futuro, o que é confirmado por [Ribeiro-Neto e Baeza-Yates, 1999].

2.7.2 Modelos Algébricos

Os modelos algébricos que serão apresentados são os modelo *indexing semantic latent* e modelo de rede neural.

2.7.2.1 Modelo indexing semantic latent

De acordo com o que foi apresentado anteriormente representar o conteúdo de um documento e de *queries* através de palavras chaves, termos indexados, pode levar a uma baixa performance de recuperação de informação, pois documentos relevantes poderiam não ser recuperados, por não estarem indexados por nenhuma das palavras chaves, bem

como documentos não relevantes a *query* poderiam ser recuperados.

[Ribeiro-Neto e Baeza-Yates, 1999], definem que, às principais razões para a ocorrência desses problemas estão relacionadas com a vagacidade (*vagueness*), combinado com o processo de recuperação de informação baseado no uso de palavras chaves.

Como base nesses problemas pode-se concluir que buscas baseadas em palavras chaves, termos indexados, não são eficientes, pois nem todos os termos conseguem representar a semântica dos documentos. Assim com a utilização de uma busca baseada na semântica dos documentos obter-se-ia uma busca mais eficiente.

Furnas et al, [apud Ribeiro-Neto e Baeza-Yates, 1999], definem que a principal concepção do modelo *indexing semantic latent* é mapear cada documento e *query* dentro de um espaço dimensional reduzido e associar com o conceito dos documentos.

Este modelo é baseado na teoria de decomposição de valor singular (*singular value decomposition),* assim apresentando uma nova estrutura teórica.

2.7.2.2 Modelo de rede neural

Rede Neural também conhecida como evolucionária, simbólica ou conexionista é uma forma de computação não algorítmica composta por sistemas que, em algum nível, apresentam uma estrutura inspirada no cérebro humano.

Para [Carvalho, Braga Antônio e Ludermir, 1998], Redes Neurais Artificiais (RNAs) são sistemas paralelos distribuídos compostos por unidades de processamento simples que computam determinadas funções matemáticas, normalmente não lineares. Tais camadas são dispostas em uma ou mais camadas e interligadas por um grande número de conexões, geralmente unidirecionais. Na maioria dos modelos estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede.

Em RNAs, é necessário uma fase de aprendizagem (treinamento), onde é fornecido um conjunto de exemplos à rede, a qual obterá as características necessárias para representar a informação fornecida.

Um modelo de rede neural para recuperação de informação proposto por [Wilkinson e Hingston, 1991], [apud Ribeiro-Neto e Baeza-Yates, 1999], é ilustrado na figura 2.4.1.

Este modelo tem sido pouco testado com grandes coleções de dados, mas tem se

mostrado com grande performance para utilização com coleções gerais.

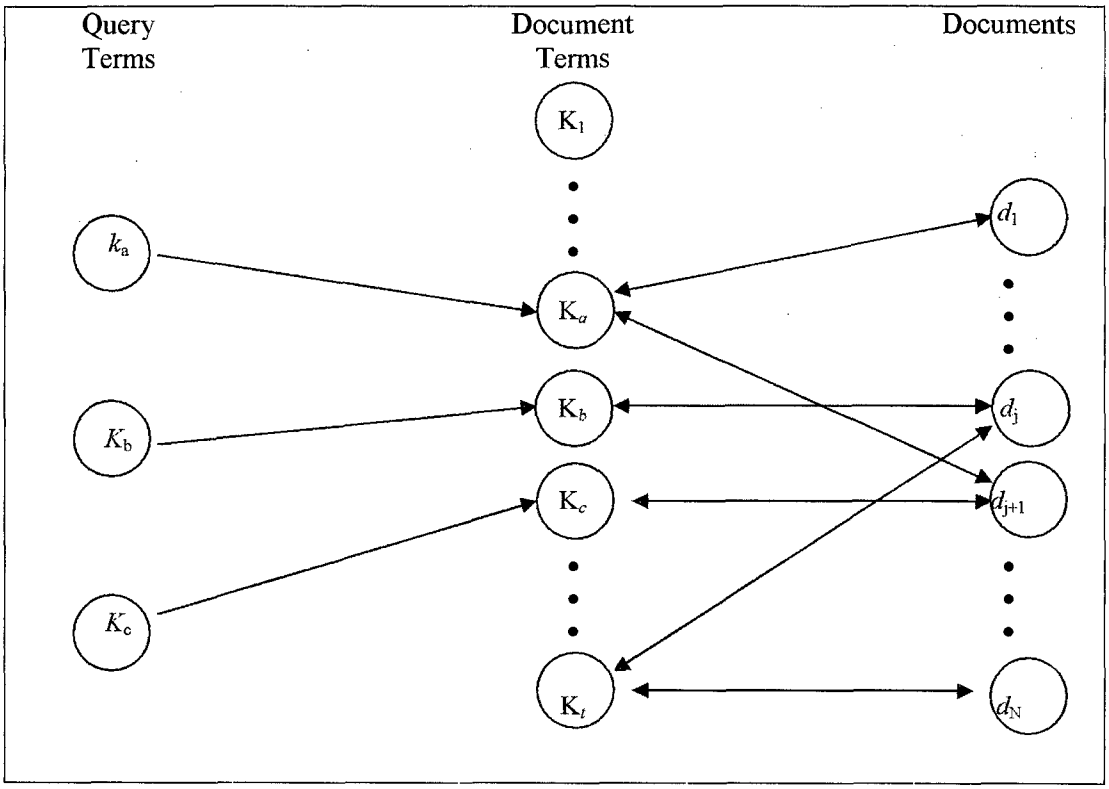


Figura 2.1- Modelo de rede neural para recuperação de informação.

2.7.3 Modelos Probabilísticos

2.7.3.1 Redes bayesianas

A utilização de teoria probabilística tem sido considerada como principal alternativa para determinar grau de relevância de documentos.

Segundo [Pearl, 1991], redes bayesianas são grafos dirigidos onde os nodos representam variáveis aleatórias e os arcos representam os relacionamentos entre essas. Os relacionamentos entre essas variáveis são influências causais cuja força é expressa por probabilidades condicionais.

De acordo com [Silva e Ribeiro-Neto, 1998], a utilização de redes bayesianas em recuperação de informação representa uma extensão do modelo probabilístico. Além de generalizar os modelos clássicos em um esquema de representação, os modelos de redes bayesianas permitem combinar características de modelos distintos para melhorar o desempenho do sistema de recuperação de informação.

Um exemplo de uma rede bayesiana é representado na figura 2.4.2.

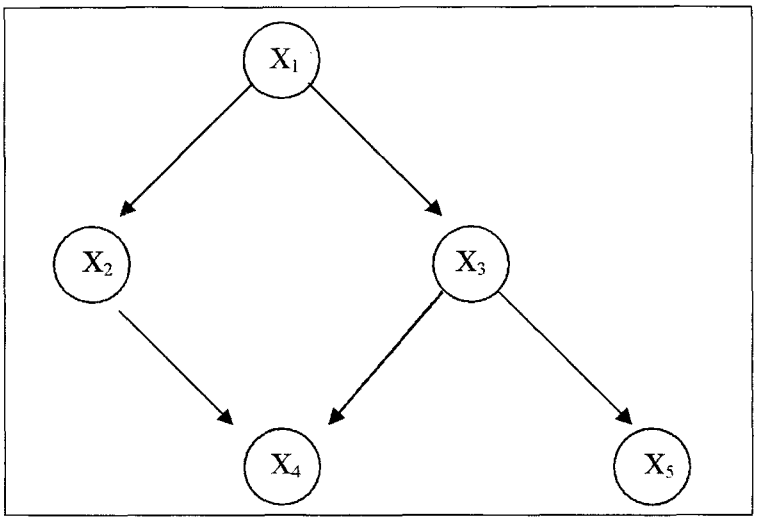


Figura 2.2- Exemplo de uma rede bayesiana

[Ribeiro-Neto e Baeza-Yates, 1999] apresentam dois modelos para recuperação de informação baseados em redes bayesianas: modelo rede de inferência (*inference network)* proposto por [Turtle and Croft, 1990 e 1991], e o modelo rede de crença (*belief network)* proposto por [Ribeiro-Neto e Muntz, 1996].

Ambos são baseados em uma visão epistemológica. Em uma visão epistemológica a probabilidade é interpretada como um grau de crença. Existe uma pequena diferença entre estes modelos que é o espaço amostral que em [Ribeiro-Neto e Muntz, 1996], é bem definido, o que não é feito em [Turtle and Croft, 1991], o que pode ser confirmado em [Silva e Ribeiro-Neto, 1998].

2.7.4 Modelos para Navegação (*browsing*)

O *browsing* é um processo no qual os usuários do sistema de recuperação não estão interessados em pesquisa baseada em *query* ou esta não pode ser usada.

Segundo (Waterworth e Chignell, 1989), este tipo de abordagem requer um modo de tomada de decisão, pois se tem que escolher a ordem, ou o caminho pelo qual se obtém as informações, o que não ocorre quanto se utiliza pesquisa baseada em *query.*

De acordo com [Ribeiro-Neto e Baeza-Yates, 1999], o termo *browsing* é o processo de procura de informações, sobrevôo ou uma maneira de navegar rapidamente pelos documentos a procura de informações relevantes. Desta forma há uma diferença entre as tarefas de pesquisa quando o alvo não é conhecido previamente e quando o alvo da pesquisa é pré-definido. Segundo [Glenn and Chignell, 1992] quando o alvo é previamente conhecido o processo é denominado de *querying*, e *browsing* é a pesquisa

cujo alvo não é previamente determinado o que é confirmado por [Ribeiro-Neto e Baeza-Yates, 1999].

A figura 2.3 ilustra a interação do usuário com o sistema de recuperação, apresentando a diferença entre recuperação de informação e *browsing*.

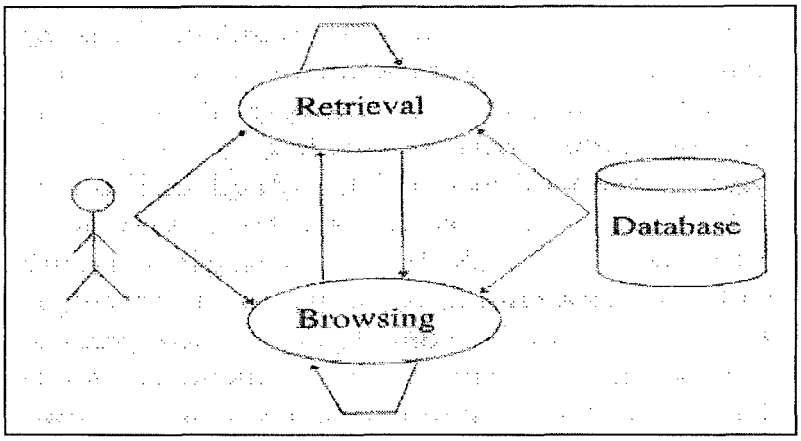


Figura 2.3- Diferença entre recuperação de informação e *browsing*.

Para [Glenn & Chignell, 1992], o *browsing* é mais natural, mas ele é menos eficiente do que um *query*, do qual se obtém um direcionamento para a informação relevante.

2.8 AVALIAÇÃO DA RECUPERAÇÃO

2.8.1 Introdução

Segundo [Ribeiro-Neto e Baeza-Yates, 1999], a avaliação da performance de sistemas de recuperação é geralmente baseada em coleções de referência e em métricas de avaliação.

Nesta seção serão apresentadas algumas das métricas de avaliação mais utilizadas bem como as coleções de referência mais importantes.

2.8.2 Métricas para Avaliação de Performance de Recuperação

Avaliação da performance de sistemas de recuperação de informação envolve além das métricas normalmente utilizadas por sistemas de recuperação de dados, tempo e espaço, a avaliação da performance de recuperação.

A avaliação da performance de recuperação nada mais é do que a precisão do conjunto resposta de uma dada *query*. Pois a partir da *query* do usuário, que por

natureza é vaga, os documentos são recuperados de acordo com sua relevância com a *query.*

A maior parte dos sistemas de recuperação de informação experimentais tem evidenciado a eficácia de recuperação de informação normalmente baseada em julgamentos de relevância do documento. Determinar a relevância de um documento é uma tarefa subjetiva e incerta, pois usuários diferentes podem ou não, designar graus de relevância diferentes para um documento, ou conjunto de documentos, recuperados a partir de uma mesma *query*. O conjunto resposta de uma dada *query* é classificado através do grau de relevância do documento com a *query,* essa classificação é o novo componente que não existe na recuperação de dados e que é parte central da problemática de avaliação de performance de recuperação.

O trabalho pioneiro nesta área foi apresentado por [Cleverdon, 1966]. [Senko 1969] e [Rijsbergen, 1979], referem que avaliação na área de sistemas de recuperação é extremamente complexa e problemática.

Sistemas de recuperação de informação podem ser avaliados baseando-se em diversos critérios, incluindo eficiência de execução, tempo gasto pelo sistema desde o processamento da entrada – *query* - até a saída, eficiência de armazenamento, eficácia de recuperação e ainda os recursos que eles oferecem a um usuário.

De acordo com [Frakes and Baeza-Yates, 1992], em sistemas de recuperação de informação muitas medidas têm sido propostas, sendo que duas delas são mais comumente utilizadas para estimar a qualidade das respostas referentes à consulta do usuário: *recall* e precisão.

Recall é definido como o quociente entre os documentos relevantes recuperados para uma dada *query,* sobre o total de documentos relevantes da coleção, recuperados ou não.

$$\textbf{\textit{Recall}} = \frac{\text{Número de documentos relevantes e recuperados}}{\text{Número total de documentos relevantes da coleção}}$$

Precisão é o quociente entre o número de documentos relevantes recuperados sobre o número total de documentos recuperados, relevantes ou não.

$$\textit{Precisão} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número de documentos recuperados}}$$

Por exemplo, supondo uma coleção qualquer de documentos na qual 80 documentos são relevantes para uma dada *query*, o sistema S retorna 60 documentos, 40 dos quais são relevantes então:

$$\begin{array}{lcll} \textit{Recall} & \Rightarrow & 40/80 & = & 50\% \\ \text{Precisão} & \Rightarrow & 40/60 & = & 67\% \end{array}$$

A figura abaixo mostra *recall* e precisão para uma dada solicitação de informação.

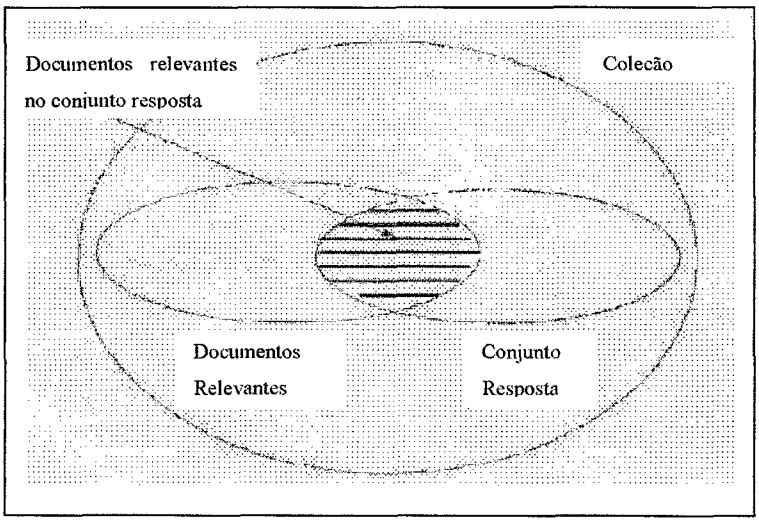


Figura 2.4- Precisão e *Recall* para uma dada *query*

Desde que freqüentemente se deseja comparar a performance de um sistema de recuperação de informação em termos de *recall* e precisão, métodos para avaliá-los simultaneamente têm sido desenvolvidos. Um método envolvendo o uso de *recall* e precisão é descrito por [Rijsbergen, 1979] e [Salton and McGill, 1983].

Precisão mede a qualidade do que foi retornado pelo sistema, isto é, a relação do que realmente é relevante com relação ao que o sistema retornou, tendo como objetivo a recuperação somente de informações relevantes, ou seja, não recuperar nada irrelevante.

Uma discussão detalhada das questões envolvidas na experimentação em recuperação de informação é apresentada em [Sparck-Jones, 1981] e [Salton e McGill,

1983].

De acordo com [Frakes e Baeza-Yates, 1992], a eficiência de execução sempre foi um assunto importante em sistemas de recuperação e a necessidade de um longo tempo de recuperação interferirá com a utilidade e vantagem do sistema.

Conforme [Ribeiro-Neto e Baeza-Yates, 1999], outras abordagens tem sido utilizadas para avaliar a performance de sistemas de recuperação tais como: média harmônica, medida E, medida orientada pelo usuário, etc.

Para [Korfhage, 1997], outras medidas podem ser interessantes tais como a satisfação, no caso de serem recuperados apenas documentos relevantes ou frustração, em caso contrário.

2.8.3 Coleções de Referência

Coleções de referência são utilizadas para avaliar a performance de sistemas de recuperação. Uma coleção de referência consiste em uma coleção de documentos, um conjunto de *queries* exemplo, um conjunto de documentos relevantes para cada *query* que são fornecidos por especialistas da área envolvida.

Nas décadas de 60, 70 e 80 não haviam parâmetros claramente definidos para avaliação da performance de sistemas de recuperação, pois grupos distintos avaliavam aspectos diferentes durante a experimentação, além de estas avaliações utilizarem coleções relativamente pequenas, o que não reflete as condições normais de utilização do sistema. Este ambiente confuso se dá pelo fato de não existir uma estrutura formal sólida com fundamentação básica para avaliação de sistemas de recuperação de informação.

A partir dos anos 90 surge a TREC (*Text REtrieval Conference*), que é uma conferência formada por organizações como o DARPA (*Defense Advanced Research Projects Agency*) e o NIST (*National Institute of Standards Technology*), além de representantes do governo americano, da indústria e da academia, que buscam integrar as novas tecnologias e técnicas desenvolvidas pela academia. O objetivo da TREC é promover conferências anuais para incentivar pesquisas na área de recuperação de informação além de padronizar procedimentos de avaliação e na realização de fóruns para comparação de resultados.

Atualmente a coleção TREC é a mais importante coleção de referência para avaliação de *queries* complexas que utilizam coleções grandes. A TREC é composta por

aproximadamente dois gigabytes de textos, cerca de um milhão de documentos e, além disso, esses números têm crescido a cada conferência.

Podem-se destacar duas outras coleções de referência que possuem importância histórica na área de recuperação de informação, são elas: CACM (coleção constituída por artigos publicados na *Communication of the ACM)* e ISI ou CISI (*Institute of Scientific Information*). Ambas fornecem um ambiente adequado para testar algoritmos para recuperação e possuem tempo de configuração pequeno, pois são coleções pequenas.

2.9 LINGUAGENS DE CONSULTA (QUERY)

2.9.1 Introdução

Uma *query* nada mais é que a expressão da necessidade de informação do usuário. Esta necessidade pode ser expressa de forma mais simplificada através de uma única palavra-chave, ou de forma mais complexa através da combinação entre várias palavras chave.

Em sistemas de recuperação de informação textuais diferentes tipos de *queries* podem ser formuladas, sendo estas totalmente dependente do modelo de recuperação utilizado.

Nesta seção objetiva-se abordar linguagens de consulta *(query)* que permitem a classificação da resposta, ou seja, linguagens para recuperação de informação como oposto à recuperação de dados, conforme mencionado anteriormente.

Serão apresentadas *queries* baseadas em palavras-chave e com operadores booleanos.

2.9.2 *Queries* Baseadas em Palavras-Chave

Esta é a forma mais comum de *query,* sendo que todos os modelos que implementam recuperação em bases textuais possibilitam pesquisa com base em uma ou várias palavras-chave.

Como resultado desde tipo de *query*, ter-se-á um conjunto de documentos que contenham a(s) palavra-chave da *query.*

O conjunto de documentos relevantes recuperados pelo sistema de recuperação pode ser classificado. Os critérios estatísticos mais comumente utilizados para

classificação de documentos com base na ocorrência de palavras no texto são: freqüência de termos e freqüência de documentos.

Freqüência de termos é o número de vezes que o termo aparece no documento e freqüência de documentos é o número de documentos em que o termo aparece.

De acordo com [Ribeiro-Neto e Baeza-Yates, 1999], *queries* baseadas em palavras-chave são consideradas *queries* básicas.

2.9.3 *Queries* de Contexto

Buscando uma maior qualidade da tarefa de recuperação muitos sistemas possuem a capacidade de pesquisar palavras em um determinado contexto.

Contexto são palavras que aparecem próximas, este fato pode indicar grande probabilidade de relevância, se estas aparecerem isoladas.

Nesta abordagem distinguem-se dois tipos de *queries:* frases e proximidade.

Uma frase é uma seqüência de caracteres, que nada mais é do que uma seqüência de q*ueries* básicas, uma ou mais palavras-chave, e apesar da utilidade deste tipo de *query* nem todos os sistemas a implementam.

Proximidade ou vizinhança por sua vez, é baseada em *queries* básicas, e mais um valor máximo permitido entre as palavras. Por exemplo: supondo que para uma dada *query Q* tenha como palavras-chave "praias Florianópolis" e o valor máximo permitido seja sete (7), assim o sistema recuperará todos o documentos que contenham as palavras-chave recuperando texto tais como: "...praia da Joaquina em Florianópolis...", "...praias do Sul do País encontram-se em Florianópolis".

2.9.4 *Queries* Booleanas

Queries booleanas são compostas pelos operadores booleanos combinada com *queries* básicas. Os operadores mais comuns utilizados são: OR, AND e BUT.

Exemplo: para uma dada *query* booleana formada por duas subexpressões, a saber, sub1 e sub2.

AND – para a *query* (sub1 AND sub2) serão recuperados todos os documentos que satisfaçam ambas as duas subexpressões;

OR – para a *query* (sub1 OR sub2) serão recuperados todos os documentos que satisfaçam uma das subexpressões, onde as duplicatas são eliminadas;

BUT - para a *query* (sub1 BUT sub2) serão recuperados todos os

documentos que satisfaçam a subexpressão sub1 e não sub2.

Assim como o modelo booleano clássico, *queries* booleanas possuem limitações, pois não permitem que o conjunto de documentos recuperados seja classificado, nem que sejam realizadas buscas parciais. Acrescenta-se a essas limitações o fato dos usuários possuírem dificuldades em expressar sua necessidade de informação através de operadores booleanos.

Devido a estas limitações um conjunto *booleano fuzzy* é proposto por [Salton, Fox and Wu, 1983] no qual a semântica dos operadores é *relaxada*, isto é ao invés do documento satisfazer totalmente o operador (AND) ele deve satisfazer pelo menos um OR. O conjunto resposta é classificado através dos documentos que possuam maior número de elementos comuns com a *query*.

O modelo probabilístico é baseado em palavras (*queries básicas*), mas também permite operações booleanas.

A tabela abaixo relaciona algumas abordagens e os tipos de operações aceitas por elas:

MODELO	QUERIES
Booleano	Palavras e operações
Vetorial	Palavras
Probabilístico	Palavras

Tabela 2.1- Relação entre abordagens e operações

2.10 OPERAÇÕES SOBRE QUERY

2.10.1 Introdução

Para muitos usuários que não apresentam conhecimento detalhado das características e do ambiente de recuperação de informação a tarefa de formular adequadamente uma *query* torna-se uma tarefa extremamente difícil. Desta forma, o usuário necessita gastar algum tempo reformulando sua *query* para obter um conjunto resposta satisfatório, isso pode ser observado na Web, em seus programas de busca.

Desse modo a *query* inicial poderia ser considerada como uma pesquisa preliminar cujo objetivo é recuperar alguma informação relevante, assim a partir destes documentos recuperados uma nova *query* seria formulada utilizando informações

extraídas destes documentos com o objetivo de melhorar o conjunto resposta.

A reformulação da *query* se dá através de dois passos básicos: expandir a *query* adicionando novos termos e reconsiderar os pesos atribuídos inicialmente aos termos. Várias abordagens utilizam estas técnicas para melhorar a *query* inicial e são classificadas em três grupos: *relevance feedback,* análise local e global.

2.10.2 Relevance Feedback

De acordo com [Ribeiro-Neto e Baeza-Yates, 1999], *Relevance Feedback* é a estratégia mais popular de reformulação de uma *query.*

Relevance Feedback é uma técnica para melhorar a efetividade da recuperação de informação baseado na avaliação de relevância dos documentos para o usuário, o que é confirmado por [Crestani e van Rijsbergen, 1997].

Salton e Buckley [apud Sparck-Jones e Willet, 1997] referem que a consulta inicial *(query)*, é uma tentativa ou hipótese e que o processo de *feedback* funciona como uma modificação controlada desta consulta inicial até a expressão que realmente representa a necessidade ou objetivo do usuário. Isto se consegue pela divisão ou quebra da consulta em consultas menores, mais simples, que podem ser melhor controladas.

Durante a execução da tarefa de recuperação de informação o usuário de um sistema de recuperação de informação propõem uma *query*, e o sistema recupera uma lista de documentos ordenados. Neste momento, o usuário avalia a relevância dos documentos recuperados, na prática, apenas os 10 ou 20 primeiros da lista, marcando aqueles que são relevantes. O sistema de recuperação de informação então modifica a *query* inicial com base nestes julgamentos de relevância. Isto é tipicamente realizado para adicionar novos termos, extraídos dos documentos relevantes, e ou para adoção de novos pesos para os termos da *query* baseados na ocorrência estatística dos documentos julgados. A partir da execução da *query* modificada uma nova lista de documentos classificados, ordenados será recuperada.

Um típico sistema de recuperação de informação que utiliza *relevance feedback* é ilustrado na figura 2.5.

A utilização de *Relevance Feedback* proporciona melhora significativa no conjunto de documentos recuperados quando comparado à recuperação baseada em *query* inicial, o que é confirmado por [Ribeiro-Neto e Baeza-Yates, 1999].

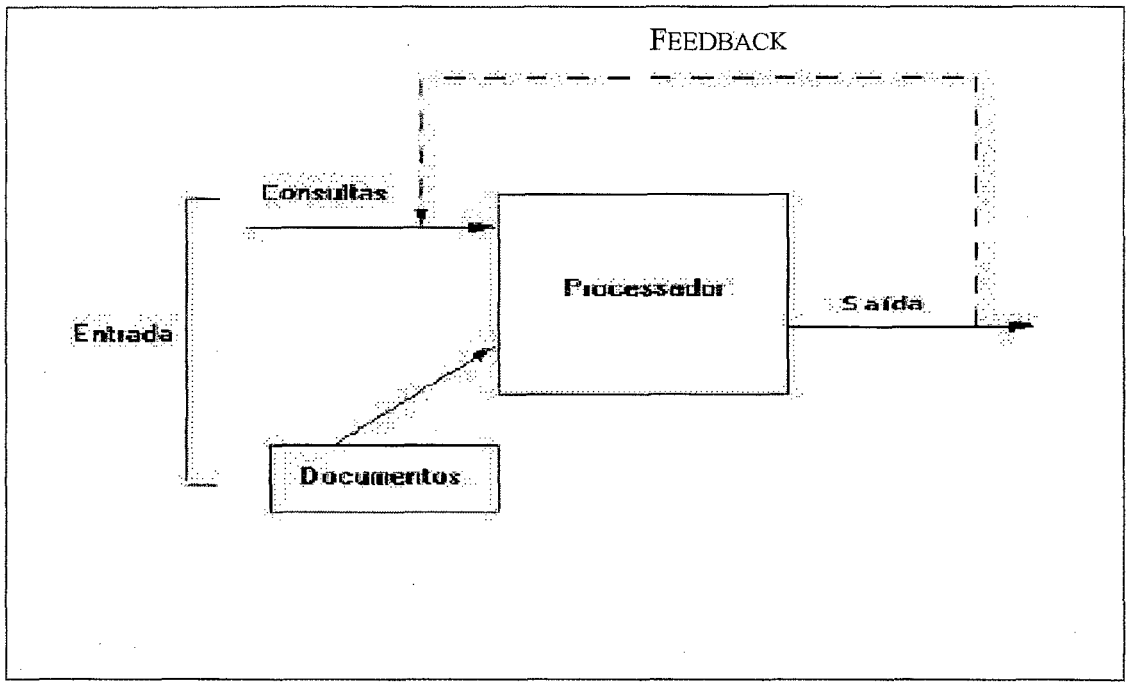


Figura 2.5- Ciclo Relevance Feedback

De acordo com [Ribeiro-Neto e Baeza-Yates, 1999] dois tipos básicos, de estratégias podem ser utilizadas para descrever o grupo de documentos relevantes automaticamente: análise local e global.

2.10.3 Análise Local

No ciclo relevance feedback o usuário classifica os documentos em dois grupos, relevantes e não relevantes, e com base nestas informações o sistema seleciona novos termos para expandir a *query* inicial, com o objetivo de recuperar mais documentos

2.10.4 Análise Global

Na análise global todos os documentos da coleção são utilizados para construção de uma estrutura global, como um dicionário que identifica o relacionamento entre os termos, de toda a coleção.

Segundo [Ribeiro-Neto e Baeza-Yates, 1999], inicialmente está abordagem foi considerada uma abordagem que não proporcionava melhora significativa na performance de recuperação. Entretanto com o advento da Internet estudos mostram que a utilização desta abordagem pode ser extremamente útil para o usuário.

A utilização de análise global combinada com a análise local pode proporcionar inúmeras vantagens para o usuário, sendo assim uma importante área para pesquisa.

2.11 OPERAÇÕES DE FILTRAGEM DO TEXTO

2.11.1 Introdução

Na linguagem natural escrita e falada, algumas palavras possuem significado mais expressivo do que outras para representar a semântica dos documentos de uma coleção. Geralmente substantivos possuem esta característica, todavia o uso de palavras-chave ou termos indexados obtidas através do pré-processamento dos documentos de uma coleção é uma prática considerada muito útil, como por exemplo, na redução do tamanho de um texto ele padroniza os documentos da coleção para simplificar a pesquisa.

2.11.2 Pré-processamento de Documentos

De acordo com [Ribeiro-Neto e Baeza-Yates, 1999], o pré-processamento de documentos pode ser dividido em cinco operações ou transformações sob o texto: análise léxica, eliminação de *stopwords*, *stemming*, escolha de termos indexados, construção de categorias de termos (*Thesauri*).

A figura 2.6 ilustra a visão lógica dos documentos durante as várias fases do pré-processamento.

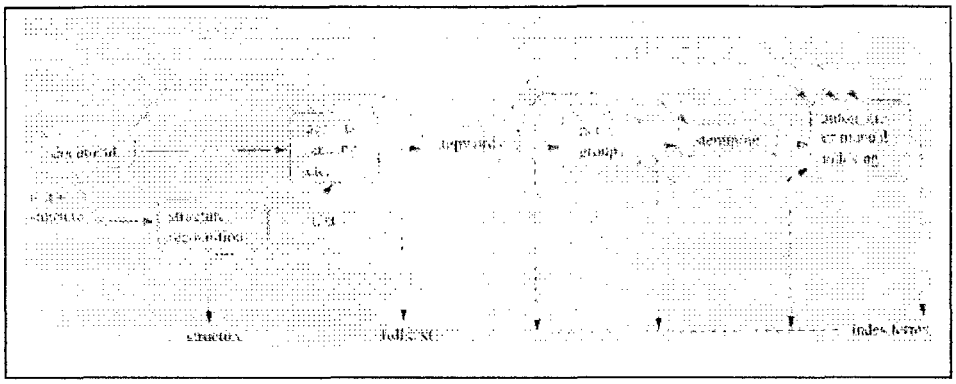


Figura 2.6- Visão lógica dos documentos durante a fase de pré-processamento

2.11.2.1 Análise léxica

A fase da análise léxica tem como objetivo principal à identificação de palavras do texto que por sua vez poderão ser utilizadas como palavras-chave. Para isso é realizado o tratamento de hífens, pontuação, números e caractere maiúsculos e minúsculos.

2.11.2.2 Eliminação de *stopwords*

Palavras que aparecem na maioria dos documentos de uma coleção não são úteis para o propósito de recuperação, tais palavras são comumente chamadas de *stopwords*, e estas normalmente não são utilizadas como palavras-chave. A lista de *stopwords* é em sua maioria formada por artigos, conjunções e preposições.

A utilização deste recurso proporciona redução considerável do tamanho da estrutura indexada, segundo [Ribeiro-Neto e Baeza-Yates, 1999], apenas com a eliminação das *stopwords* pode-se chegar à redução de aproximadamente 40% desta.

2.11.2.3 Stemming

Ao formular uma *query* o usuário utiliza uma palavra, mas apenas uma das variações sintáticas desta palavra tais como, plural, variações verbais, entre outras, estará presente no documento relevante. Estas variações sintáticas impedem uma recuperação dita perfeita.

Este problema pode ser solucionado parcialmente através da substituição das palavras por seu radical (*stem*).

Stemming tem como o objetivo eliminar prefixos e sufixos (afixos), bem como

permitir a recuperação de variações sintáticas dos termos da *query* e dos documentos da coleção. Um exemplo típico da utilização deste recurso é apresentado por [Ribeiro-Neto e Baeza-Yates, 1999], no qual a palavra *connect* é o radical das variações *connected, connecting, connection* e *connections*.

Segundo [Frakes and Baeza-Yates, 1992], a utilização de *stem* proporciona melhora na performance de recuperação, além de reduzir o tamanho da estrutura indexada.

Em [Frakes and Baeza-Yates, 1992], encontra-se um estudo comparativo entre oito diferentes trabalhos sobre os benefícios potenciais da utilização de *stemming* para proporcionar melhor eficácia da recuperação.

[Ribeiro-Neto e Baeza-Yates, 1999], referem que não há consenso na literatura atual sobre os benefícios do *stemming* na melhora da performance de recuperação.

2.11.2.4 Termos indexados (palavras-chave)

Segundo [Baeza-Yates, 1996], ferramentas de recuperação de informação geralmente trabalham com técnicas de indexação capazes de indicar e acessar mais rapidamente documentos de uma coleção de documentos.

Podem-se destacar três tipos de indexação: indexação tradicional, indexação *full-text* e indexação por *tags*.

Indexação tradicional

Indexação tradicional é aquela, onde o usuário determina os termos que caracterizam os documentos, e estes farão parte do índice de busca.

Este tipo de indexação encontra-se na maioria dos sistemas bibliotecários, e a criação de *Thesaurus* (índice hierárquico) é uma das preocupações desta área. Contudo, a intervenção humana é o maior problema deste tipo de indexação, pois pode gerar falhas no índice, além de ser uma atividade difícil e que consome muito tempo.

Indexação *full-text*

Na indexação *full-text* ou indexação do texto todos os termos que compõem o documento fazem parte do índice, sem estruturas hierárquicas entre estes. As ferramentas indexam automaticamente todas as palavras do documento, gerando índices volumosos. Para redução destes índices podem-se utilizar técnicas como árvores TRIE, árvores PAT, listas invertidas, etc.

De acordo com [Baeza-Yates, 1996], podem-se utilizar filtros nesta técnica de indexação para evitar a indexação de termos indesejáveis tais como *stopwords*, o que

proporciona uma redução no tamanho deste índices.

Segundo [Ribeiro-Neto e Baeza-Yates, 1999], este tipo de abordagem é adotado por alguns dos programas de busca da Web.

Indexação por *tags*

Indexação por *tags* ou por partes do texto procura indexar somente as partes relevantes do documento. Para tanto, as ferramentas automatizadas deverão analisar cada documento procurando por marcas, *tags*, que identificam estas partes mais importantes.

Para a indexação por *tags*, a maioria dos trabalhos adota o uso de gramáticas, *parsers*, expressões regulares e autômatos finitos para a definição e identificação das marcas.

O maior problema das técnicas de indexação por *tags* é que os tipos de documentos devem ter sido analisados previamente, ou seja deve-se conhecer de antemão que tipo de conteúdo compõe os documentos e quais as marcas que identificam as partes relevantes de um documento. Então as ferramentas acabam sendo muito específicas para determinados tipos de informação e documentos.

2.11.2.5 Thesaurus

Thesaurus são estruturas léxicas que organizam e relacionam semanticamente as palavras de uma linguagem.

Em sua forma mais simples o thesaurus consiste de:

1) Uma lista de palavras importantes de uma determinada área do conhecimento; e
2) Para cada palavra desta lista, um conjunto de palavras relacionadas, geralmente derivadas de sinônimos.

Segundo [Ribeiro-Neto e Baeza-Yates, 1999], a motivação para construção de um *thesaurus* esta baseada na utilização de um vocabulário preciso que sirva no contexto de recuperação de informação, para coordenar a indexação e recuperação de documentos.

[Ribeiro-Neto e Baeza-Yates, 1999], apresentam as vantagens da utilização de *thesaurus* preciso que incluam: normalização dos conceitos de indexação, identificação das palavras-chave indexadas com significado semântico claro, entre outras.

2.12 ESTRATÉGIAS DE BUSCA

De acordo com [Rijsbergen, 1979], todas as estratégias de busca são baseadas na comparação entre a *query* e os documentos armazenados. A forma mais simples de busca é a busca seqüencial.

Os mais importantes algoritmos de busca seqüencial são: algoritmo de força bruta, o algoritmo de Knuth-Morris-Pratt, o algoritmo de Karp-Rabin, diferentes variações do algoritmo de Boyer-Moore, entre outros. Um estudo detalhado deste algoritmos pode ser encontrado em [Frakes and Baeza-Yates, 1992] e [Ribeiro-Neto e Baeza-Yates, 1999].

2.13 INDEXAÇÃO

O objetivo principal de técnicas de indexação é construir estruturas de dados eficientes sobre o texto, chamadas índices, que proporcionem pesquisas rápidas nos textos de uma coleção. Índices são coleções de termos indexados que apresentam ponteiros para lugares onde informações relacionadas a estes podem ser encontradas.

De acordo com [Ribeiro-Neto e Baeza-Yates, 1999], a construção e o gerenciamento de índices não são recomendados para coleções de dados grandes ou semi-estáticas, isto é, que sofrem atualizações em um determinado espaço de tempo.

Segundo [Frakes and Baeza-Yates, 1992], a implementação de buscas indexadas é extremamente simples e eficiente para uso em pequenas e médias coleções de dados, o que é confirmado por [Ribeiro-Neto e Baeza-Yates, 1999].

Existem muitas classes de índices baseados em diferentes abordagens de recuperação. [Ribeiro-Neto e Baeza-Yates, 1999] referem que três das principais técnicas são: arquivos invertidos, *suffix arrays* e *signature files.*

Suffix arrays é uma estrutura pouco comum por apresentar muitos detalhes de implementação além de grande dificuldade no seu gerenciamento, o que torna sua utilização restrita a algumas áreas tais como a genética.

De acordo com [Ribeiro-Neto e Baeza-Yates, 1999], *signature files* foi uma técnica bastante popular nos anos 80, sendo que atualmente arquivos invertidos apresentam uma melhor performance.

Desta forma arquivos invertidos é atualmente a técnica de indexação mais

adequada, pois além de apresentar melhor performance que outras técnicas possui maior flexibilidade no tratamento de novos tipos de *queries*.

O conceito de arquivos invertidos é apresentado por [Frakes e Baeza-Yates, 1992], da seguinte forma: suponha um conjunto de documentos, para cada documento é designados uma lista de palavras-chave ou atributos, com pesos de relevância opcionais associados com cada palavra-chave, atributo. Um arquivo invertido é então uma lista classificada ou índice de palavras-chave, com cada palavra-chave tendo ligações com os documentos contendo aquela palavra-chave.

A representação deste exemplo pode ser observada na figura 2.7.

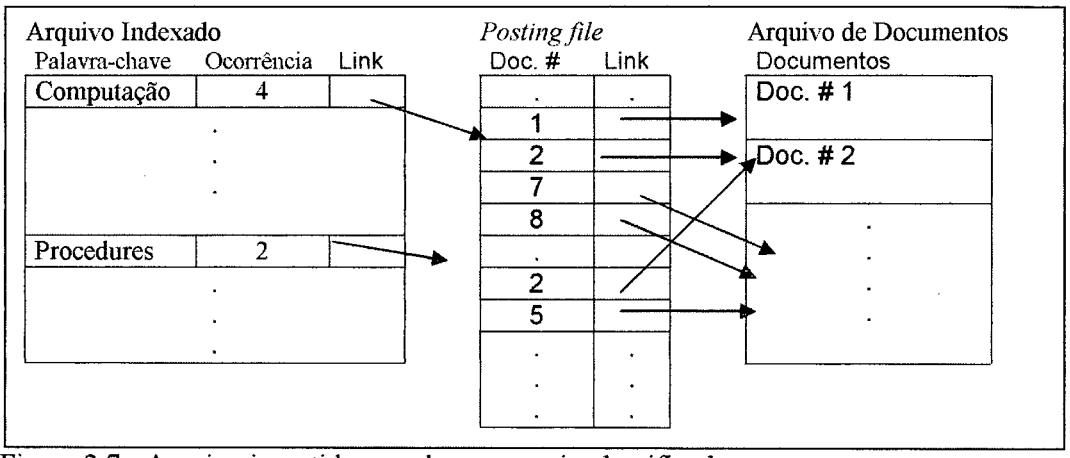


Figura 2.7- Arquivo invertido usando um arranjo classificado.

[Frakes and Baeza-Yates, 1992] refere que, quase todos os tipos de índices são baseados em alguma forma de árvore ou *hashing*. Dentre as exceções a esta colocação destacam-se: *clustering* e DAWG (*Direct Acyclic Word Graph*).

Clustering é uma operação sob uma coleção de dados que busca agrupar em classes os documentos similares.

Segundo [Blumer et al., 1985], DAWG é baseado na teoria de autômatos finitos e tem como objetivo representar todas às sub-palavras possíveis do texto, utilizando uma quantia linear de espaço.

3 LÓGICA PARACONSISTENTE

3.1 INTRODUÇÃO

[Da Silva Filho e Abe, 1999], referem que a lógica paraconsistente foi desenvolvida de forma autônoma e independente em 1948, por Stanislaw Jaskowski, polonês e por Netwon C. A. Da Costa, brasileiro.

A lógica clássica considera apenas dois estados lógicos, o verdadeiro e o falso. Portanto a lógica clássica não faz consideração sobre valores entre o verdadeiro e falso. Considerando-se o mundo em que vivemos, pode-se apresentar inúmeras situações cotidianas onde o uso da lógica clássica torna-se ineficaz, pois a utilização da lógica clássica ou tradicional no tratamento de situações que envolvam inconsistência, ambigüidade, paradoxos ou vagacidade (*vagueness*) apresenta inúmeras deficiências.

Diversas pesquisas foram e estão sendo desenvolvidas em busca de aplicações para as lógicas alternativas da lógica clássica, denominadas Lógicas não Clássicas. As lógicas não clássicas tratam de maneira mais adequada situações que apresentem indefinições, paradoxos e inconsistências.

Conforme [Da Silva Filho e Abe, 1999], lógica paraconsistente é uma lógica não clássica que aceita e trata situações que envolvam contradições. Uma lógica é dita paraconsistente se esta permite que uma conclusão possa ser obtida de premissas contraditórias como, por exemplo: a contradição $(p \wedge \sim p)$ é ao mesmo tempo verdadeira e falsa, como oposto a simplesmente falsa.

Lógica paraconsistente permite esta e outras espécies de contradições e inconsistências.

Existem várias classes de lógicas paraconsistentes. A lógica paraconsistente anotada, que é uma destas classes, será apresentada na seção 3.3.

3.2 MOTIVAÇÃO E APLICAÇÃO DA LÓGICA PARACONSISTENTE

Várias pesquisas sobre lógica paraconsistente estão sendo realizadas nos últimos anos, como pode ser observado em trabalhos publicados em [Laptec'2000, 2000].

Segundo [Da Costa e Marconi, 1987], pode-se encontrar inúmeras aplicações da

Lógica Paraconsistente nas mais variadas áreas de conhecimento, tais como: matemática, filosofia, ciências e tecnologia. Pesquisas mais recentes incluem ainda: lógica epsitêmica, física, psicanálise, medicina, hardwares, entre outras.

Pode-se destacar várias pesquisas em importantes sub-áreas da ciência da computação, tais como: Inteligência Artificial e Computacional, Robótica, Engenharia de Produção, o que é confirmado por [Da Costa, Da Silva Filho, Abe, et al., 1999] e [Da Silva Filho e Abe, 1999] e pode ser observado em [Laptec'2000, 2000].

Como exemplo da aplicação nas áreas de Robótica, Inteligência Artificial e Computacional pode-se salientar o projeto do Robô Emmy, um robô móvel autônomo baseado totalmente na teoria de lógica paraconsistente, [Da Silva Filho, Abe, Mário, et al., 1999].

3.3 MODELAGEM PARACONSISTENTE PARA CONHECIMENTO HUMANO

[Da Costa e Subrahmanian, 1989], referem que situações que apresentam inconsistências e ambigüidades são comuns na descrição do mundo real, e a lógica clássica não é capaz de apresentar bons resultados, ou sua aplicação de forma direta fica impossibilitada, o que é confirmado por [Da Costa, Da Silva Filho, Abe, et al., 1999].

[Da Costa, Da Silva Filho, Abe, et al., 1999], apresentam exemplos de situações cotidianas que apresentam inconsistências tais como: uma reunião de condomínio para decidir sobre uma reforma; um administrador que deve promover um de seus funcionários. Esses exemplos têm, como objetivo demonstrar que a aplicação de lógica paraconsistente para modelar conhecimento por meio de evidências obtêm resultados muito semelhantes aos resultados obtidos através de raciocínio humano.

3.4 LÓGICA PARACONSISTENTE ANOTADA COM DOIS VALORES – $LPA2_V$

3.4.1 Considerações Iniciais

Segundo [Da Costa, Da Silva Filho, Abe, et al., 1999], as lógicas anotadas são uma das classes de lógicas paraconsistentes.

Os primeiros estudos sintáticos e semânticos da lógica anotada foram desenvolvidos concomitantemente por [Da Costa, Subrahmanian e Vago, 1991] e [Da Costa, Abe e Subrahmanian, 1991].

De acordo com [Da Silva Filho e Abe, 1999], a lógica paraconsistente anotada pode ser representada por um diagrama de Hasse, conforme a figura 3.1.

Na lógica paraconsistente anotada as informações são obtidas na forma de graus de crença relativos a uma dada proposição. Nesta lógica os valores de crença e descrença variam dentro do intervalo [0,1], generalizando a lógica bi-valorada, como pode ser observado na figura 3.1(b), onde para cada vértice é atribuído uma variável que pode ser considerada um estado lógico, figura 3.1(a).

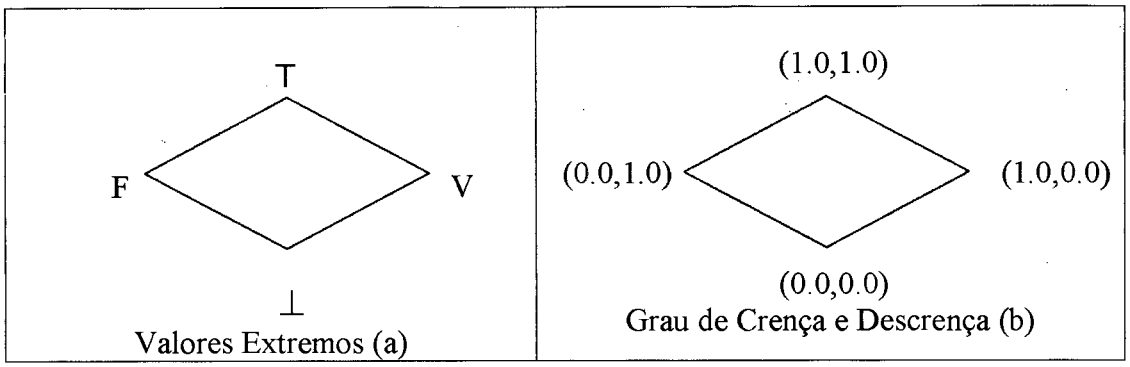


Figura 3.1 – Diagrama representativo da Lógica Paraconsistente Anotada.

Com base na figura 3.1, é possível determinar um relacionamento entre os estados lógicos e os valores de crença e descrença da seguinte forma:

T	⇔	(1.0, 1.0)	⇔	Inconsistente
V	⇔	(1.0, 0.0)	⇔	Verdadeiro
F	⇔	(0.0, 1.0)	⇔	Falso
⊥	⇔	(0.0, 0.0)	⇔	Indeterminado

Tabela 3.1- Relacionamento entre estados lógicos e graus de crença e descrença

3.4.2 Lógica Paraconsistente Anotada e Graus de Crença e Descrença

O reticulado que traz dois valores na anotação foi proposto por [Da Costa, 1990]. A aplicação de tal retículo em conjunto com a lógica paraconsistente permite melhora significativa no desempenho de programas computacionais, o que é confirmado por [Da Costa, Da Silva Filho, Abe, et al., 1999].

Os estados lógicos apresentados (figura 3.1a) são encontrados com base em dois valores de anotação (figura 3.1b), os quais são descritos por um par (μ_1, μ_2) sendo que o

par (μ_1, μ_2) representam respectivamente, o grau de crença e descrença atribuído a proposição.

[Da Costa, Da Silva Filho, Abe, et al., 1999], referem que a idéia epistemológica intuitiva de associação de uma preposição *p* a uma anotação (μ_1, μ_2) significa que o grau de crença em *p* é de no máximo μ_1, e o grau de descrença da preposição *p* é de no máximo μ_2.

Em tal diagrama (1.0, 0.0) indica "crença total", (0.0, 1.0) indica "descrença total", (1.0, 1.0) indica "crenças totalmente inconsistentes" e (0.0, 0.0) indica "ausência total de crença. Exemplificando: seja a preposição *p* = " Hoje não irá chover. " Temos:

Se anotarmos (1.0, 0.0), pode-se deduzir que: hoje não irá chover com crença total, ou seja, crê-se totalmente que hoje não irá chover;

Se anotarmos (0.0, 1,0) pode-se deduzir que: hoje não irá chover com descrença total, crê-se totalmente que hoje irá chover;

Se anotarmos (1.0, 1.0) pode-se deduzir que: hoje não irá chover com crenças totalmente inconsistentes. Isso pode acontecer se um Instituto de meteorologia prever que hoje irá chover e outro prever ao contrário.

Se anotarmos (0.0, 0.0) pode-se deduzir que: hoje não irá chover com ausência total de crença. Isso pode acontecer se nenhum Instituto de metereologia for capaz de prever o tempo para hoje.

Aplicações da lógica paraconsistente que fazem uso do diagrama *quatro* consideram quatro situações (figura 3.1a), das quais duas não são tratadas pela lógica clássica, a inconsistência e a indefinição. Apesar destas aplicações apresentarem diferenças significativas em relação à lógica clássica não se obtêm um bom grau de precisão, por serem considerados apenas valores extremos e constantes.

Anand, Subrahmanian e Flog apud [Da Costa, Da Silva Filho, Abe, et al., 1999], propõem um reticulado associado à Lógica Paraconsistente Anotada que pode aumentar a precisão, onde duas novas anotações são inseridas. (Figura 3.2).

Conforme [Da Silva Filho e Abe, 1999], na utilização prática de lógica paraconsistente anotada as entradas são os graus de crença e descrença da proposição, enquanto que os estados lógicos são as saídas, obtidas através da análise paraconsistente, como é ilustrado na figura 3.3.

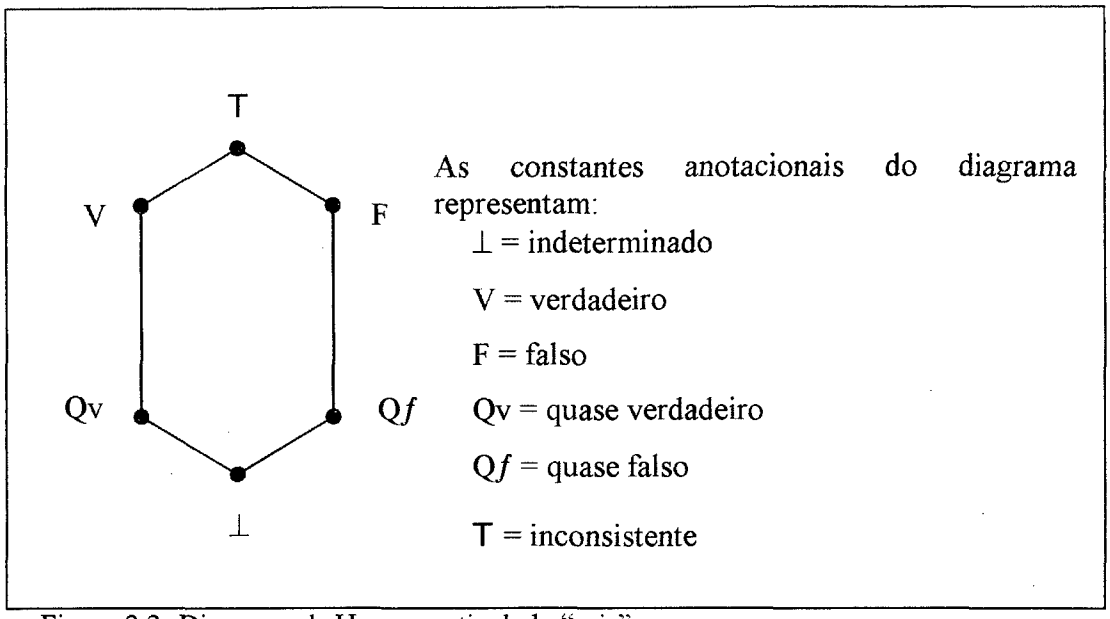


Figura 3.2- Diagrama de Hasse – reticulado "seis"

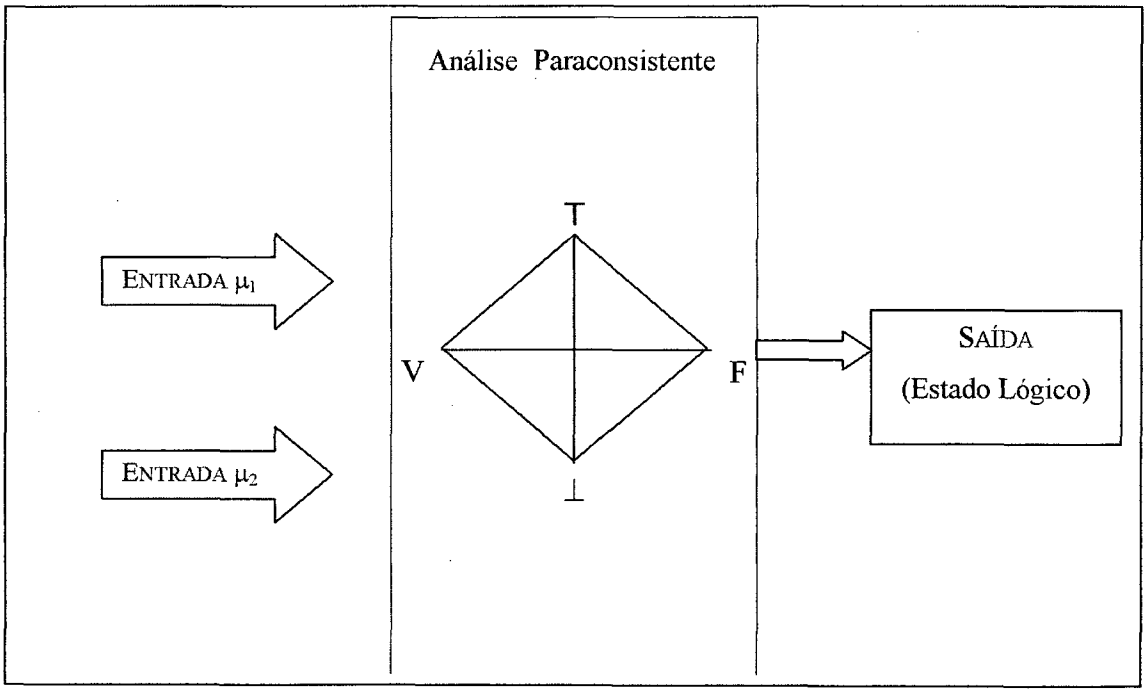


Figura 3.3- Sistema básico de análise paraconsistente

3.4.3 Análise da Lógica Paraconsistente Anotada de Anotação com Dois Valores no Quadrado Unitário do Plano Cartesiano - QUPC

De acordo com [Da Silva Filho e Abe, 1999], a análise paraconsistente dos graus de crença e descrença é obtida através da representação do diagrama em um Quadrado Unitário no Plano Cartesiano (QUPC), figura 3.4, onde os graus de crença (certeza) ficam representados no eixo x (μ_1), e os graus de descrença (incerteza) ficam expostos no eixo y (μ_2).

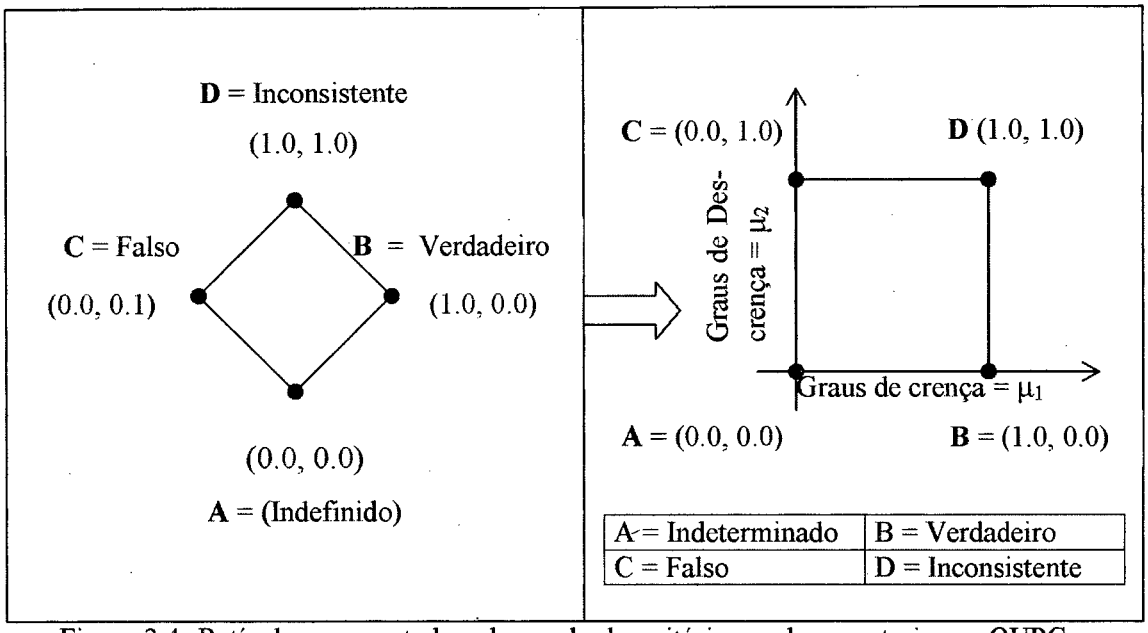


Figura 3.4- Retículo representado pelo quadrado unitário no plano cartesiano - QUPC

Conforme pode ser observado na figura 3.4, os valores dos graus de crença e descrença são binários. [Subrahmanian, 1987] refere que os valores de crença (μ_1) e de descrença (μ_2) podem ser iguais ou diferentes. Desta forma o estado lógico obtido como resultado no QUPC pode ser indeterminado, inconsistente, falso ou verdadeiro o que pode ser observado na figura 3.4. Estes resultados são denominados estados extremos.

Através de pontos interpolados no diagrama é possível identificar o grau de crença e descrença da proposição. Cada um destes pontos corresponde a um estado lógico como saída, assim dependendo das necessidades da aplicação desejada pode-se aumentar o número de pontos situados dentro do QUPC, que resultará como conseqüência num aumento proporcional a quantidade de pontos delimitados nos estados lógicos como saída, além dos valores extremos já determinados.

De acordo com [Da Costa, Da Silva Filho, Abe, et al., 1999], para aumentar a

precisão da aplicação os valores dos graus de crença e descrença podem ser ternários e independentes. Com a utilização dos valores ternários para graus de crença e descrença interpolados no QUPC, obterão-se-á cinco novos pontos no retículo, os quais por sua vez definirão cinco novos estados lógicos resultantes (tabela 3.2), denominados estados lógicos não extremos.

ESTADOS NÃO EXTREMOS	SÍMBOLO
Verdadeiro, tendendo ao inconsistente	$V_{\rightarrow}T$
Verdadeiro, tendendo ao indeterminado	$V_{\rightarrow}\perp$
Falso, tendendo ao inconsistente	$F_{\rightarrow}T$
Falso, tendendo ao indeterminado	$F_{\rightarrow}\perp$
Quase verdadeiro	Q-v

Tabela 3.2 – Estados lógicos não extremos

A representação de todos os estados lógicos extremos e não extremos é apresentada na figura 3.5.

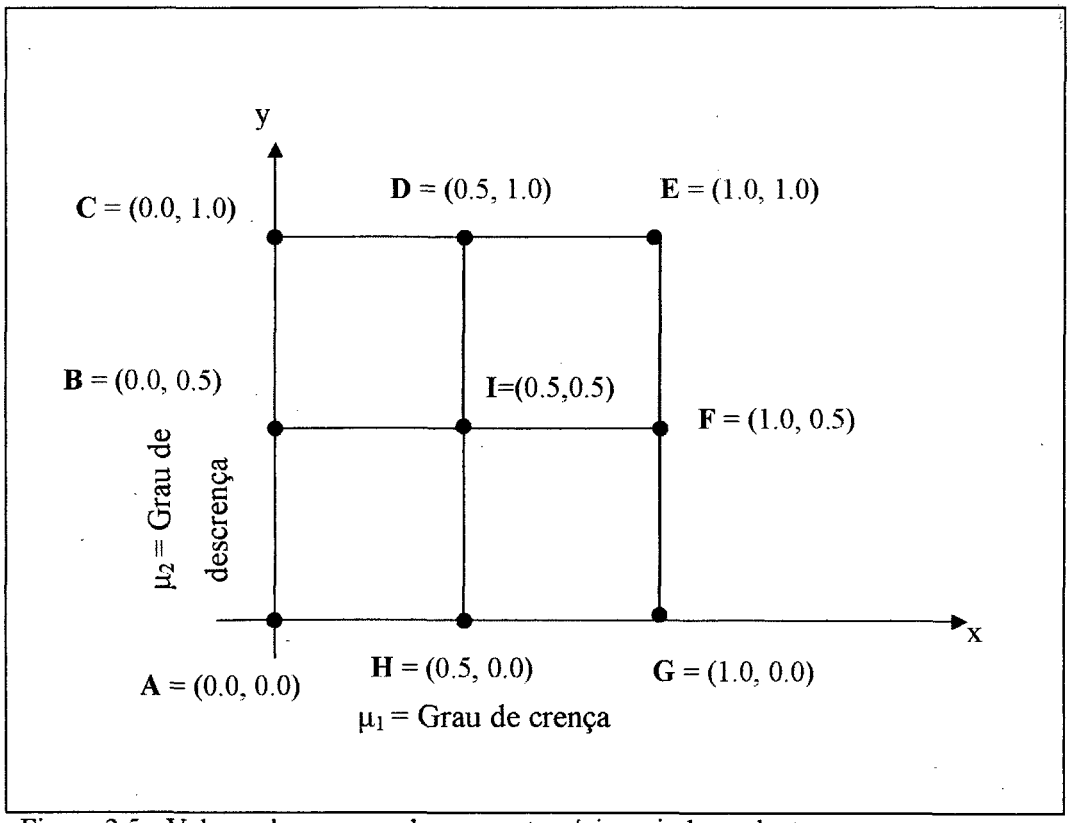


Figura 3.5 - Valores de crença e descrença ternários e independentes

A tabela 3.3 relaciona as variáveis da figura 3.5 com seus respectivos estados

lógicos.

VARIÁVEL		ESTADO LÓGICO
A	⇔	Indeterminado
B	⇔	Falso, tendendo ao indeterminado
C	⇔	Falso
D	⇔	Falso, tendendo ao inconsistente
E	⇔	Inconsistente
F	⇔	Verdadeiro, tendendo ao inconsistente
G	⇔	Verdadeiro
H	⇔	Verdadeiro, tendendo ao indeterminado
I	⇔	Quase verdadeiro

Tabela 3.3 – Relação entre estados lógicos extremos e não extremos e os pontos no QUPC

[Da Costa, Da Silva Filho, Abe, et al., 1999], referem que através de um par ordenado (μ_1, μ_2) pode-se calcular o grau de contradição (G_{ct}) e o grau de certeza (G_c) da proposição.

O grau de contradição (G_{ct}) é definido como o valor que representa no diagrama à distância entre dois estados lógicos extremos, definidos como inconsistente e indeterminado. O grau de contradição (G_{ct}) é composto pelo grau de indeterminação (G_{id}) e pelo grau de inconsistência (G_{it}). Desta forma o grau de contradição encontra-se no intervalo fechado [-1,1]. A figura 3.6 ilustra o eixo dos valores de contradição.

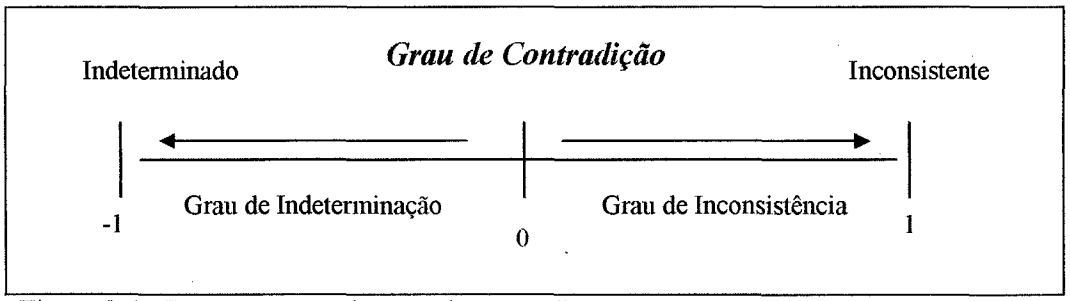


Figura 3.6 - Representação do grau de contradição

O grau de certeza (G_c) é definido como o valor que representa no diagrama a distância entre dois estados extremos, denominados falso e verdadeiro. O grau de certeza (G_c) é composto pelo grau de verdade (G_v) e pelo grau de falsidade (G_f), os quais possuem valores no intervalo fechado [-1,1]. A figura 3.7 representa o eixo dos valores de certeza.

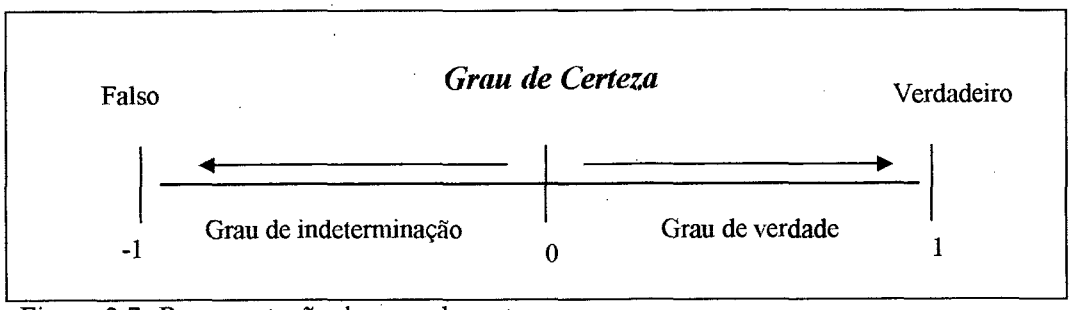


Figura 3.7- Representação do grau de certeza

Para encontrar os graus de certeza e contradição deve-se considerar dois segmentos de reta definidos sob o QUPC: $\overline{BD}$, que será chamado de linha perfeitamente definida (figura 3.8a) e o segmento $\overline{AC}$ que será chamado de linha perfeitamente indefinida. (figura 3.8 b).

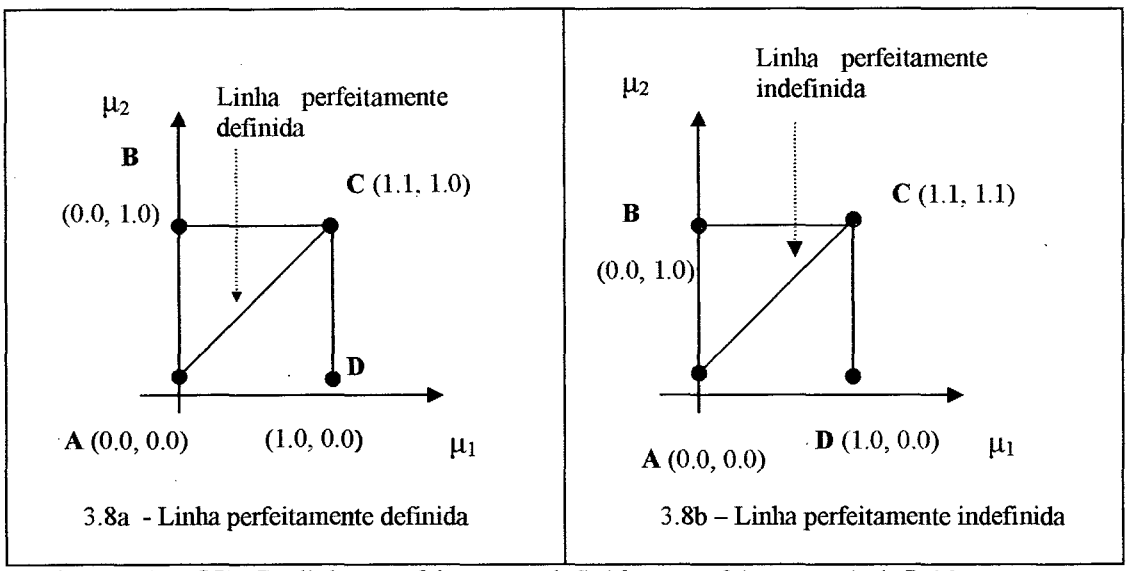


Figura 3.8 - QUPC e linhas perfeitamente definidas e perfeitamente indefinidas

Cada par (μ_1, μ_2) gera um único ponto situado dentro do QUPC. A figura 3.8a será utilizada para calcular o grau de contradição e a figura 3.8b será utilizada para encontrar o grau de certeza da preposição.

Considerando a figura 3.8a, quando um ponto estiver situado abaixo da linha perfeitamente definida, o resultado da equação que envolve os graus de crença e descrença será negativo, por outro lado se estiver situado acima desta o resultado será positivo.

Para definição das equações do grau de contradição (G_{ct}) considera-se a figura

3.8a. Desta forma quando o grau de contradição (G_{ct}) for maior ou igual a zero será denominado de grau de inconsistência (G_{it}). Quando o grau de contradição for menor que zero será denominado grau de inconsistência (G_{id}). As equações que definem o grau de inconsistência são apresentadas na tabela 3.4.

O grau de inconsistência apresenta valores no intervalo fechado [0,1], e o grau de indeterminação no intervalo [-1,0], na linha perfeitamente definida ambos têm valor igual a zero. Portanto:

$$0 \leq G_{it} \leq 1 \qquad \text{e} \qquad -1 \leq G_{id} < 0$$

Considerando a figura 3.8b, quando um ponto estiver situado abaixo da linha perfeitamente indefinida o resultado da diferença entre os graus de crença e descrença será positivo, e esse resultado é o valor do grau de verdade (G_v). Por outro lado se o ponto estiver acima da linha perfeitamente indefinida o resultado da diferença entre os graus de crença e descrença será negativo, define-se esse resultado como grau de falsidade (G_f), da preposição. As equações que definem tanto grau de verdade quando falsidade são apresentadas na tabela 3.4.

Na linha perfeitamente indefinida ambos graus de verdade e falsidade possuem valor igual a zero. Enquanto o grau de verdade possui valor no intervalo [0,1] o grau de falsidade apresenta-se no intervalo [-1,0]. Desta forma:

$$0 \leq G_v \leq 1 \qquad \text{e} \qquad -1 \leq G_f < 0$$

A partir do par ordenado (μ_1, μ_2) são encontrados os valores dos graus de verdade (G_v), de falsidade (G_f), de inconsistência (G_{it}) e de indeterminação (G_{id}), conforme tabela 3.4 a seguir:

G_{it}	=	$\mu_1 + \mu_2 - 1$	se e somente se:	$(\mu_1 + \mu_2) \geq 1$
G_{id}	=	$\mu_1 + \mu_2 - 1$	se e somente se:	$(\mu_1 + \mu_2) < 1$
G_v	=	$\mu_1 - \mu_2$	se e somente se:	$\mu_1 \geq \mu_2$
G_f	=	$\mu_1 - \mu_2$	se e somente se:	$\mu_1 < \mu_2$

Tabela 3.4 – Equações do grau de certeza e contradição

Convêm ressaltar que os valores de falsidade e indeterminação terão seus valores considerados em módulo, conforme pode ser observado na figura 3.9 que apresenta marcação dos valores máximos para os graus de certeza (verdade e falsidade) e

contradição (indeterminação e inconsistência).

Segundo [Da Costa, Da Silva Filho, Abe, et al., 1999], a partir dessa interpretação da LPA2v e fazendo análise do QUPC, pode-se fazer a seguinte afirmação:

> Um ponto gerado pelos graus de crença μ_1 e de descrença μ_2 está situado dentro do quadrado unitário do plano cartesiano e pode ser descrito através de valores do grau de inconsistência G_{it}, do grau de indeterminação G_{id}, do grau de verdade G_v, e do grau de falsidade G_f.

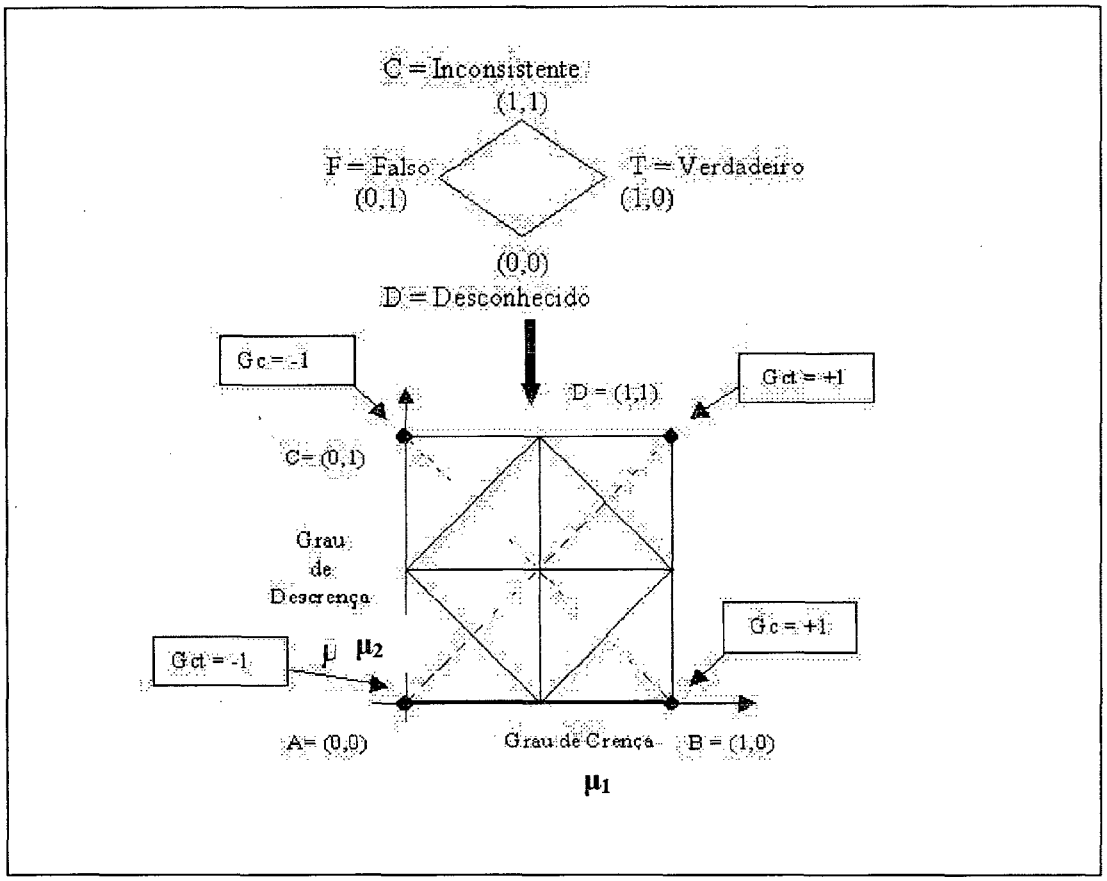


Figura 3.9 - Reticulado da Lógica Paraconsistente Anotada, representado num quadrado

De acordo com [Da Silva Filho e Abe, 1999], na prática um sistema paraconsistente funciona da seguinte forma:

1. Se o grau de contradição for muito elevado, significa que não há certeza ainda quanto à decisão, desta forma é necessário obter-se novas evidências (proposições);
2. Se houver um alto grau de certeza a conclusão pode ser gerada, desde que haja um baixo grau de incerteza.

Através da análise paraconsistente um único estado lógico será atingido ao término de cada análise. A análise paraconsistente é traduzida por um algoritmo chamado de "Para-Analisador" (descrito na seção 3.5), através da análise dos valores de crença e descrença obter-se-á os graus de certeza e contradição da proposição, conforme ilustrado na figura 3.10.

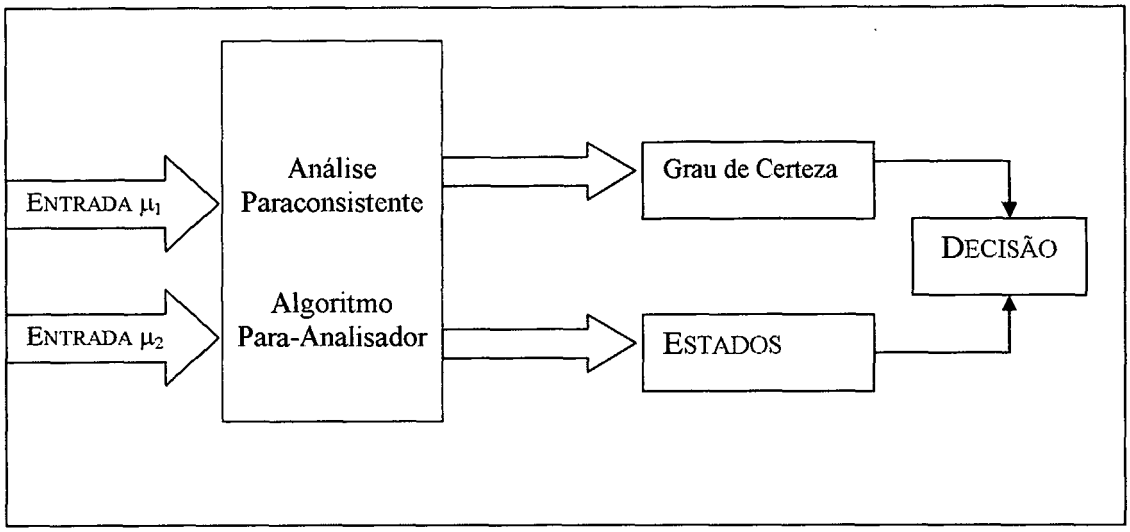


Figura 3.10- Análise paraconsistente com algoritmo para-analisador

Convêm ressaltar que o grau de contradição (G_{ct}), e o grau de certeza (G_c) estão intimamente ligadas no reticulado representativo da LPA_V, desta forma não é possível determinar o estado resultante considerando estes valores de forma individual.

Para melhor representar essa dependência à figura 3.11 apresenta os graus de G_{ct} e G_c em dois eixos.

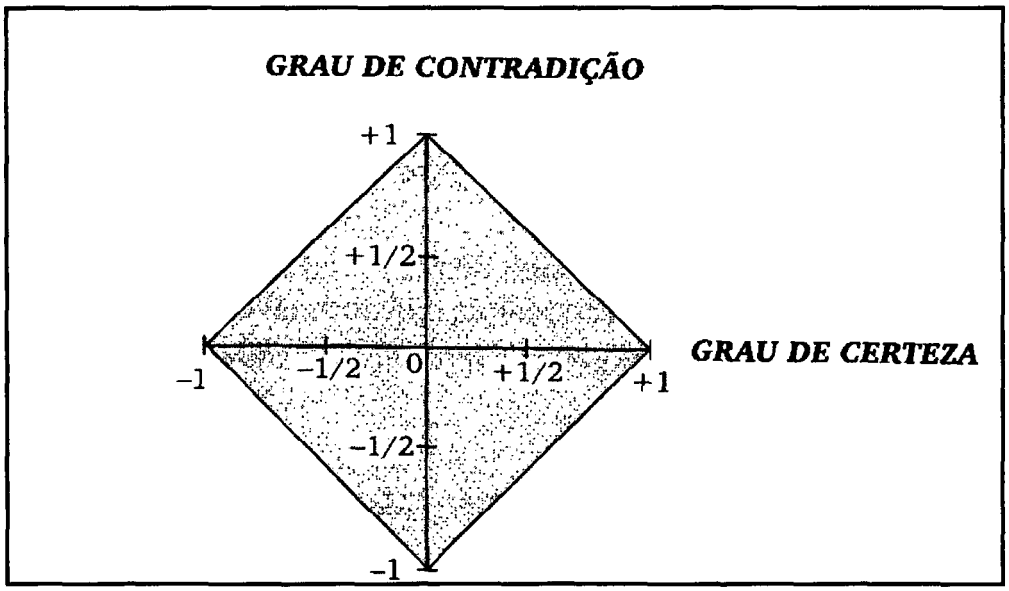


Figura 3.11- Representação dos graus de certeza e de contradição inter-relacionados

3.5 .ALGORITMO "PARA-ANALISADOR"

*/ Definição dos valores */

V_{scc} = C_1 */ Definição do valor superior de controle de certeza */

V_{icc} = C_2 */ Definição do valor inferior de controle de certeza */

V_{scct} = C_3 */ Definição do valor superior de controle de contradição */

V_{icct} = C_4 */ Definição do valor inferior de controle de contradição */

*/ Variáveis de entrada */

μ_1

μ_2

*/ Variáveis de saída */

S_1 = Saída discreta

S_{2a} = Saída analógica

S_{2b} = Saída analógica

*/ Expressões matemáticas */

sendo: $0 \leq \mu_1 \leq 1$ e $0 \leq \mu_2 \leq 1$

$Gct = \mu_1 + \mu_2 - 1$

$Gc = \mu_1 - \mu_2$

*/ Determinação dos estados extremos */

Se $Gc \geq C_1$ então $S_1 = V$

Se $Gc \leq C_2$ então $S_1 = F$

Se $Gct \geq C_3$ então $S_1 = \top$

Se $Gct \leq C_4$ então $S_1 = \bot$

*/ Determinação dos estados não extremos */

Para $0 \leq Gc < C_1$ e $0 \leq Gct < C_3$

Se $Gc \geq Gct$ então $S_1 = Qv \rightarrow \top$

Se $Gc < Gct$ então $S_1 = \top \rightarrow v$

Para $0 \leq Gc < C_1$ e $C_4 < Gct \leq 0$

Se $Gc \geq |Gct|$ então $S_1 = Qv \rightarrow \perp$

Se $Gc < |Gct|$ então $S_1 = \perp \rightarrow v$

Para $C_2 < Gc \leq 0$ e $C_4 < Gct \leq 0$

Se $|Gc| \geq |Gct|$ então $S_1 = Qf \rightarrow \perp$

Se $|Gc| < |Gct|$ então $S_1 = \perp \rightarrow f$

Para $C_2 < Gc \leq 0$ e $0 \leq Gct < C_3$

Se $|Gc| \geq Gct$ então $S_1 = Qf \rightarrow \top$

Se $|Gc| < Gct$ então $S_1 = \top \rightarrow f$

ct 2a

c 2b

De acordo com [Da Costa, Da Silva Filho, Abe, et al., 1999] há várias possibilidades para as aplicações do Algoritmo Para-Analisador, tais como:

- controlador de processo industrial;
- controle de áreas de automação;
- analisador em sistemas especialistas;
- inteligência artificial
- robótica.

4 THESAURUS

4.1 INTRODUÇÃO

Segundo [Ribeiro-Neto e Baeza-Yates, 1999], *thesaurus* são conjuntos de índices, na forma de frases ou palavras e mais um conjunto de relações *(links)* entre esses índices. Índices do tipo thesaurus talvez sejam os mais conhecidos e usados. São tipos específicos de dicionários e seguem a técnica do vocabulário controlado.

Thesaurus são estruturas que possuem grande valor para Sistemas de Recuperação de Informação (SRI), pois estes proporcionam um vocabulário preciso e controlado, através do qual é possível coordenar o processo de indexação e recuperação de documentos, o que é confirmado por [Frakes e Baeza-Yates, 1992].

A função mais comumente utilizada de um *thesaurus* é a de selecionar termos para representar documentos. Conforme referido anteriormente, o termo documento é aqui empregado de forma geral, podendo este se referir a revistas, artigos, jornais, entre outros.

Thesaurus pode ser utilizado no processo de definição da melhorar sua estratégia de busca. Por exemplo: o usuário de um sistema de recuperação de informação expressa sua necessidade de informação através de uma *query*, caso para essa dada *query* o sistema não recupere nenhum documento, o *thesaurus* pode ser utilizado para expandir os termos da *query* inicial através das relações existentes entre seus índices, conforme pode ser observado em [MINKER, et al., 1972]. Por outro lado, se esta *query* retorna-se muitos documentos o *thesaurus* poderia ser utilizado para sugerir termos mais específicos, este recurso a utilizado por *thesaurus* on-line.

O uso de *thesaurus* em recuperação de informação envolve construção automática, desenvolvimento de interface com o usuário, mecanismos e arquitetura de recuperação.

Esta seção esta organizada da seguinte forma: a primeira seção 4.2 apresenta a definição de dicionário, a seção 4.3 descreve os recursos de um *thesaurus* e a seção 4.4 comenta as formas de construção de *thesaurus*. As seções 4.5 e 4.6 apresentam duas das principais abordagens para construção automática de *thesaurus*.

4.2 DICIONÁRIO

Dicionário é um bom exemplo do uso de vocabulário controlado e servem também para a classificação de termos em categorias.

Segundo [GUTHRIE, et al., 1996], o objetivo de um dicionário é dar informações sobre palavras tais como: etimologia, pronúncia, morfologia, sintaxe e significados, provendo assim conhecimento sobre a linguagem e também sobre o mundo.

Índices do tipo *thesaurus* (tipos específicos de dicionários), talvez sejam os mais conhecidos e usados.

4.3 CARACTERÍSTICAS DE *THESAURUS*

4.3.1 Nível de Coordenação

Coordenação refere-se à construção de frases a partir de termos individuais. *Thesaurus* reconhece duas formas distintas de coordenação: pré-coordenação e pós-coordenação.

Thesaurus pré-coordenados são aqueles que contém frases, possibilitando assim a utilização destas no processo de indexação ou recuperação. Por outro lado, *thesaurus* pós-coordenados não contém frases, mas estas são construídas durante o processo de busca.

Definir qual das duas abordagens é melhor é uma tarefa complicada, pois ambas possuem vantagens e desvantagens das quais podem-se destacar as seguintes: a pré-coordenação possui um vocabulário mais preciso o que proporciona redução de ambigüidade na indexação e pesquisa. Porém possui a desvantagem de que o usuário necessita ter conhecimento prévio das regras envolvidas para construção das frases.

Porém *thesaurus* permite coordenação híbrida, ou seja, a utilização de ambos: frases e palavras simples (únicas). Tal prática é comum em *thesaurus* construídos manualmente.

Segundo [Frakes e Baeza-Yates, 1992], existem diferenças quanto ao nível de pré-coordenação, pois algum thesaurus pode fazer uso de duas ou três palavras como frases, enquanto outros podem fazer uso de um número bem maior.

Pós-coordenação o usuário não necessita conhecer as regras de construção da

frase, ou seja, a ordem com que as palavras devem aparecer. Combinações de frases podem ser construídas durante o processo de pesquisa.

[Salton e McGill, 1983], referem que pós-coordenação possui falhas na precisão, pois é difícil diferencias frases como, por exemplo: "*Venetian blind*" - "blind Venetian" e "library school" – "school – library".

Segundo [Frakes e Baeza-Yates, 1992], esse problema acontece devido à ausência total de regras, tornando possível a recuperação de vários documentos irrelevantes a necessidade de informação do usuário.

Pré-coordenação é comum em *thesaurus* construídos manualmente e pós-coordenação é geralmente utilizada por *thesaurus* construídos automaticamente, embora a construção de frase automaticamente seja extremamente difícil, o que é confirmado por [Frakes e Baeza-Yates, 1992].

4.3.2 Relacionamento entre os Termos

[Aitchison e Gilchrist, 1972], classificam os relacionamentos entre os termos em três categorias: relacionamentos equivalentes, relacionamentos hierárquicos e não hierárquicos.

Relacionamentos equivalentes incluem sinônimos e quase-sinônimos. No processo de recuperação de informação o significado de quase-sinônimos são aqueles termos que podem ser considerados sinônimos, por exemplo, os termos "genéticos" e "hereditários".

Relacionamentos hierárquicos possuem relação gênero-espécie, tal como "cachorro" e "lhasa apso" ou "gato" e "persa".

Relacionamentos não-hierárquicos identificam conceitos relacionados aos termos, tais como: atributos, partes, entre outras.

Uma classificação alternativa pode ser encontrada em [Wang, et. al., 1985], que define cincol categorias de relacionamentos. De acordo com [Frakes e Baeza-Yates, 1992], tal classificação é similar a classificação apresentada por [Aitchison e Gilchrist, 1972] e além de que algumas destas categorias não apresentam significado muito claro para recuperação de informação, embora alguns trabalhos como [Fox, et al., 1988] se baseiam nesta classificação.

4.3.3 Especificação do Vocabulário

Podemos definir conjunto de termos organizados de forma hierarquizada, com o objetivo de possibilitar a recuperação de informações, reduzindo assim consideravelmente a diversidade de terminologia.

Uma base de dados que utiliza um vocabulário controlado possibilita o desenvolvimento de estratégias de busca (recuperação) a partir daquelas palavras-chave listadas no *thesaurus*.

Um conjunto muito amplo de termos gera ambigüidades e problemas com uso de sinônimos, ou seja, um mesmo termo para representar vários conceitos e um mesmo conceito representado por vários termos.

[Korfhage, 1997], caracteriza isto como falta de consistência nos termos usados na indexação, seja esta manual ou automática.

Uma das maneiras de reduzir tais problemas é fixando o conjunto de termos que podem ser usados, isto é, vocabulário controlado.

De acordo com [Frakes e Baeza-Yates, 1992], o objetivo principal é identificar os termos (palavras ou frases) que melhor representam os documentos da coleção para estes fazerem parte do vocabulário do *thesaurus*, da referida coleção.

Entretanto, quando se limita o conjunto de termos, outros problemas podem surgir. Um deles se refere ao uso de nomes próprios, os quais não podem ser definidos em pré-coordenação.

4.3.4 Normalização do Vocabulário

Normalização do vocabulário consiste de um extenso conjunto de regras que gerencia as entradas dos termos no *thesaurus*. Na construção de *thesaurus* manualmente é dada uma ênfase maior nesta fase.

Alguns exemplos de tais regras são demonstrados abaixo:

Apenas substantivos (nomes) são aceitos como termos;

Limitar os número de adjetivos que pode ser usado em frases;

Usar termos apenas no singular; entre outras.

A construção manual de um *thesaurus* envolve um número bastante significativo de regras para cada um dos termos da estrutura. Desta forma o usuário deve ter conhecimento desta regras de normalização para utilizar tal sistema, tornando-se assim a

maior desvantagem da normalização. Por outro lado tem-se um vocabulário bastante consistente.

Já as regras de normalização de *thesaurus* construídos automaticamente são mais simples e envolvem *stop list* e *stemming* como será demonstrado na seção 4.5.4.

4.4 CONSTRUÇÃO DE *THESAURUS*

A construção de thesaurus ocorrer de duas formas: manual ou automática.

4.4.1 Construção Manual de *Thesaurus*

De acordo com [Frakes e Baeza-Yates, 1992], a construção manual de thesaurus é uma arte e uma ciência.

O primeiro passo para construção manual de *thesaurus* é a definição da área (assunto). A definição destes limites inclui, a identificação da área principal, subáreas e a definição de grau de importância de cada uma destas.

Após a definição do assunto da área principal, bem com das respectivas subáreas a coleção de termos para cada subárea pode ser iniciada com o auxilio de diversos recurso tais como: índices, enciclopédias, jornais, resumos e usuários do sistema. O próximo passo é analisar o vocabulário a fim de detectar sinônimos, termos relacionados, entre outros. O próximo passo é organizar os termos em uma estrutura hierárquica, através de seus relacionamentos.

Dentre às abordagens para construção de *thesaurus* manualmente, pode-se destacar duas: a primeira possui objetivos gerais, e é baseada em palavras. Podem-se apresentar os seguintes exemplos: Roget's, [Roget, 1988] e Word Net, [MILLER, et al., 1993]. Apesar de conter correlações entre os termos tais como, antônimos e sinônimos esse sistemas são raramente utilizadas em sistemas de recuperação de informação.

A segunda é orientada a recuperação de informação e é baseada em frases. Pode-se destacar INSPEC, LCSH (*Library of Congress Subject Headings*).

[Frakes e Baeza-Yates, 1992], referem que a maior dificuldade encontrada na construção de *thesaurus* manualmente esta em seu alto custo de construção bem como na dificuldade de atualização, o que é confirmado por [Ribeiro-Neto e Baeza-Yates, 1999].

Segundo [Frakes e Baeza-Yates, 1992] as desvantagens apresentadas no

desenvolvimento de um *thesaurus* manualmente torna-se o maior incentivo para o desenvolvimento de *thesaurus* de forma automática.

4.4.2 Construção Automática de *Thesaurus*

Thesaurus construídos de forma automática é geralmente dependente do domínio, ou seja, depende da base de dados (coleção de documentos) na qual ele será utilizado.

As primeiras pesquisas relacionadas à construção de *thesaurus* automaticamente foram desenvolvidas por [Spark e Needham, 1968], [Salton, 1968], [Sparck e Jackson, 1970], [Rijsbergen, Harper, Porter, 1981] e [Salton, Buckley e Yu, 1983].

Thesaurus construídos automaticamente são baseados na co-ocorrência de informação e no julgamento de relevância e geralmente são utilizados para estimar a probabilidade que o termo do *thesaurus* possui de apresentar similaridade com os termos da *query*.

No trabalho de [Qui e Frei, 1993] foi construída uma matriz de similaridade de *term-vs-term*, baseada na forma como os termos de uma coleção são indexados. Neste estudo um método probabilístico é usado para estimar a probabilidade de um termo ser semelhante (similar) com uma dada *query* no modelo espaço vetorial ou simplesmente modelo vetorial.

[Croft e Jing, MA01003] referem que os experimentos de [Salton, 1968], [Spark e Needham, 1968], [Sparck e Jackson, 1970], [Rijsbergen, Harper, Porter, 1981] e [Salton, Buckley e Yu, 1983], fazem utilização de classificação de termos automaticamente sem o julgamento de relevância ou *feedback* não produz qualquer melhora significante, o que é confirmado posteriormente por [Minker, Wilson, Zimmerman, 1972].

Contudo é importante destacar a existência de uma vasta literatura de princípios e metodologia, porém apenas uma pequena parte esta relacionada a construção automática de *thesaurus*.

4.5 *THESAURUS* CONSTRUÍDOS A PARTIR DE TEXTOS

De acordo com [Frakes e Baeza-Yates, 1992], o processo de construção pode ser dividido em três etapas: 1ª - construção do vocabulário; 2ª - determinação da similaridade entre os termos e o 3ª - organizar o vocabulário.

4.5.1 Construção do Vocabulário

A construção de um vocabulário tem como objetivo maior, identificar os termos mais relevantes (significativos) para estes constituírem o vocabulário do *thesaurus* de uma determinada coleção de documentos.

Para construção de um vocabulário é necessário inicialmente identificar a coleção de documentos, pois como referido anteriormente, *thesaurus* construídos automaticamente são dependentes do domínio. Depois de identificada a coleção de documentos (área – assunto), é necessário determinar as especificações do *thesaurus*. Dependendo do nível de especificação requerido pode-se, por exemplo, selecionar os termos a partir do título ou resumo do documento para um alto nível de especificação. Para um baixo nível de especificações é possível utilizar-se todo o documento, se este estiver disponível.

Depois de definidos os termos que constituem inicialmente o vocabulário, este esta pronto para a próxima etapa que é a normalização.

Segundo [Frakes e Baeza-Yates, 1992], o processo de normalização do vocabulário divide-se basicamente em duas fases: eliminação de palavras comuns (*stoplist)* e *stem*.

A eliminação de palavras comuns (triviais) tais como preposições e conjunções e o processo mais comum de normalização, e consiste da construção de uma lista com estas palavras como pode ser observado em [Fox, 1990].

No processo de normalização padrão o próximo passo é a aplicação de *stem* no vocabulário. O processo de *stemming* consiste de transformar a palavra em seu radical de origem, conforme referido anteriormente (seção 2.11.2.3).

4.5.1.1 Avaliação e seleção de *stem*

Existem uma série de métodos estatísticos para avaliação da relevância (peso) do termo. [Frakes e Baeza-Yates, 1992] destacam três métodos: 1° - seleção de termos baseado na freqüência de ocorrência, 2° - seleção de termos baseado na discriminação de valor e 3° - seleção de termos baseados no modelo Poisson. O programa *select.c* (Anexo – 2) inclui rotinas para os três métodos.

O método de seleção pela freqüência de ocorrência foi baseado no trabalho de Luhn's. Este método é o mais antigo e a idéia base é que cada termo pode ser enquadrado em três categorias de freqüência com relação à coleção de documentos. São

elas: alta, média e baixa freqüência.

De acordo com [Frakes e Baeza-Yates, 1992], os termos que se enquadram em média freqüência são os melhores para o processo de indexação e pesquisa.

Já [Salton e McGill, 1983], recomendam a criação de classes termos com baixa freqüência, por outro lado [Frakes e Baeza-Yates, 1992] referem que não há evidencias de como criar essas classes de forma automática.

Segundo [Salton e Yang, 1973], o método de seleção pela discriminação de valor (DV) é uma medida que corresponde ao grau de capacidade que cada termo possui para distinguir ou discriminar documentos dentro de uma coleção de documentos.

Por último, a seleção através do método *Poisson* é uma distribuição aleatória (randômica), discreta que pode ser usada como modelo. Tal método faz parte da família de modelos *Poisson* que inclui: [BOOKSTEIN e SWANSON, 1974] [HARTER, 1975] [SRINIVASAN, 1990].

4.5.1.2 Construção de frases

Conforme referido anteriormente, a construção de frases é dependente do nível de coordenação definido no caso se o nível de coordenação for pré-coordenação a construção de frases é possível.

A construção de frases é normalmente executada para reduzir a freqüência de termos que possuem alta freqüência proporcionando assim que estes termos possuam um aumento de seu valor para a recuperação, pois termos que possuem média freqüência são os que proporcionam melhor qualidade para recuperação.

Os métodos para construção de frase de [Salton e McGill, 1983] e de [Choueka, 1988] são apresentados a seguir.

Procedimento de Salton e McGill

Este procedimento estatístico é uma alternativa aos métodos sintático e semântico para a construção de frases.

De forma geral, este procedimento baseia-se em dos critérios gerais. O primeiro define que as palavras, que compõem a frase, podem aparecer com uma certa freqüência em um determinado contexto, como em uma sentença. De uma forma mais rigorosa, as palavras podem aparecer a uma certa distância uma da outra.

O segundo critério define que as palavras que constituem a frase podem aparecer com uma certa freqüência (freqüência alta) devido ao fato destas representarem a idéia

central do texto. Em [Frakes e Baeza-Yates, 1992] encontra-se um algoritmo, motivado por este segundo critério, chamado de *select.c*.

Procedimento Choueka

O algoritmo proposto por [Choueka, 1988] é estatístico e combinatório e requer uma coleção de documentos grande (pelo menos um milhão de itens) para que o mesmo seja eficiente. O trabalho do referido autor propõe uma nova abordagem para identificação de expressões coloquiais, o mesmo refere que o significado de uma frase não pode ser derivado apenas das palavras que a compõem.

A principal diferença entre os dois procedimentos esta no número de termos permitidos para constituir a frase.

4.5.2 Determinação da Similaridade

Após a definição de um vocabulário adequado e construídas frases, se isto for preciso, a próxima etapa é definir a similaridade entre os pares de termos.

Há várias formas de medidas de similaridade disponível na literatura. [McGill, et al, 1979] apresenta um estudo comparativo entre vários destes métodos. Em [Frakes e Baeza-Yates, 1992], o programa *select.c* apresenta duas rotinas de cálculo de similaridade: coseno e dado.

Essas medidas podem ser usadas para avaliar a similaridade ou associações entre os termos de uma coleção de documentos.

4.5.3 Organização do Vocabulário

O último passo é definir a estrutura do vocabulário que geralmente define um arranjo hierárquico das classes de termos.

E possível encontrar na literatura vários programas de agrupamento (*clustering*), mas uma algoritmo padrão geralmente aceita todos os pairwise correspondentes para uma coleção de documentos.

4.6 FUSÃO (MERGING) DE *THESAURUS* EXISTENTES

A fusão (*merge*) de *thesaurus* é extremamente útil quando diferentes *thesauri* estão disponíveis para algum assunto. Em [Frakes e Baeza-Yates, 1992] o programa *merge.c* executa a fusão entre diferentes hierarquias.

4.7 BREVE DESCRIÇÃO DOS PROGRAMAS

Será apresentada aqui uma breve descrição dos programas descritos nesta seção. Os referidos programa encontram-se nos sites dos autores do livro [Frakes e Baeza, 1992] que são: http://frakes.cs.vt.edu ou http://dcc.uchile.cl/~rbaeza/.

4.7.1 Programa *Select.c*

Esse programa contém várias rotinas para seleção do critério a ser utilizado no desenvolvimento do vocabulário do *thesaurus*.

De acordo com [Frakes e Baeza, 1992] o referido programa requer dois arquivos de entrada, um normal e outro invertido, ambos com as mesmas informações, diferenciando-se apenas no arranjo, como pode ser observado na figura 4.1.

2 math 2.0	mellitus 1 1.0
2 logic 1.0	logic 2 1.0
1 diabetes 2.0	diabetes 1 2.0
1 mellitus 1.0	math 2 2.0
3 math 1.0	math 3 1.0

Figura 4.1 – Arquivos de Entrada – invertido e normal

4.7.2 Programa *Hierarchy.c*

Esse programa pode executar duas funções. A primeira, dado um relacionamento hierárquico entre um conjunto de termos, este grava esse relacionamentos num arquivo invertido, como demonstrado na figura 4.1. O segundo pode gerar a estrutura hierárquica automaticamente usando o algoritmo Rada´s [FORSYTH e RADA, 1986].

4.7.3 Programa *Merge.c*

Esse programa contém rotinas para executar dois tipos de funções de fusão (*merge*).

5 RECUPERAÇÃO DE INFORMAÇÃO EM BASES TEXTUAIS: UMA ABORDAGEM PARACONSISTENTE

5.1 INTRODUÇÃO

A presente dissertação investiga o uso da lógica paraconsistente para recuperação de informação em bases textuais.

A tarefa de recuperação de informação em bases textuais é baseada em recuperação de informação não estruturada. Tais informações podem incluir texto, imagens, áudio etc. O propósito desta dissertação irá se restringir a sistemas de informação baseado em texto. Na tarefa de recuperação de informação padrão, documentos são pré-definidos e o sistema recuperará documentos que satisfaçam as necessidades de informação do usuário (*query*).

Em uma aplicação típica, a coleção de documentos pode ter vários gigabytes de tamanho (exemplo TREC) e ainda conter milhares de documentos. Em uma coleção desse tamanho, geralmente apenas uma centena de documentos, ou menos, poderá ser relevante para uma *query* especifica.

A grande disparidade na cardinalidade do conjunto de documentos relevantes versus o conjunto de documentos não relevantes torna o problema completamente diferente de muitas tarefas de classificação e influencia como sistemas de recuperação são projetados e como eles são avaliados.

Uma coleção de documentos pode ser indexada por um conjunto de atributos. Em um sistema de recupéração baseado em texto, atributos podem incluir palavras, frases, itens de vocabulário controlados manualmente, entre outros. O foco desta dissertação será em métodos automáticos e atributos determinados manualmente.

Este capítulo está organizado da seguinte forma: na seção 2 descreve dados textuais e na seção 3 são descritos alguns modelos de recuperação de informação em bases textuais. Na seção 4 é apresentado o quadrado unitário do plano cartesiano de resolução 12^1 que será utilizado para avaliar a relevância do termo no modelo aqui proposto. Em seguida na seção 5 comenta os benefícios da utilização de vocabulário controlado e conjunto com linguagem natural. E na seção 6 é apresentado o modelo para recuperação de informação baseado em lógica paraconsistente.

5.2 DADOS TEXTUAIS

Dados textuais fazem parte da categoria de dados não-estruturados. Os textos contêm as informações, mas suas representações (os dados) não são claras, nem explícitas, muito menos precisas (pode haver ambigüidades).

O armazenamento de textos pode-se dar por meio de textos livres (escritos em alguma linguagem natural sem organização nenhuma), de textos semi-estruturados (contendo alguma estrutura), de textos em linguagem restrita (um subconjunto da linguagem natural com formatos específicos), de dicionários e até mesmo através de campos tipo texto em sistemas estruturados (por exemplo, resumo de um livro em um sistema bibliográfico).

5.3 MODELOS DE RECUPERAÇÃO DE INFORMAÇÃO EM BASES TEXTUAIS

Atualmente, em vários meios observa-se a grande necessidade de armazenamento de dados textuais. Sobre esses armazenamentos há a necessidade de recuperação de informação. O crescimento constante do volume dos dados armazenados em bases textuais acarreta o crescimento proporcional da dificuldade na manipulação destes, o que torna cada vez mais complexa a tarefa de recuperação de informação.

Vários modelos têm sido propostos para trabalhar com recuperação de dados em bases textuais. Entre eles estão: o modelo vetorial proposto por Salton apud [Lima, Laender, Ribeiro-Neto, 1998], o modelo hierárquico proposto por [Lima, Laender e Ribeiro-Neto, 1998], o modelo baseado em rede bayesiana proposto por [Silva e Ribeiro-Neto, 1998] e o modelo baseado em filtros proposto por [Mendonça, Silva e Ribeiro-Neto, 1998].

Bases textuais geralmente são grandes, e dependendo do domínio tendem a crescer com tempo.

Para [Mendonça, Silva e Ribeiro-Neto, 1998], a recuperação de informação em bases textuais é normalmente modelada por uma coleção de documentos indexados por palavras chaves. Nesta abordagem, o usuário especifica sua necessidade de informação através de uma *query* (conjunto de palavras chaves), objetivando recuperar todos aqueles documentos que possuam alguma relação com sua necessidade de informação.

A definição de quais termos representará melhor a *query* e os documentos são tarefas que apresentam grande complexidade, pois a maioria das bases textuais utiliza linguagem natural a qual é repleta de inconsistências, paradoxos e incertezas.

A razão pela qual será utilizada lógica paraconsistente como uma abordagem para recuperação de informação em bases textuais decorre da facilidade proporcionada por esta no tratamento de incertezas, paradoxos, inconsistências e vagacidade (*vagueness)*.

5.4 QUADRADO UNITÁRIO DO PLANO CARTESIANO DE RESOLUÇÃO 12^1

Conforme referido anteriormente o aumento do número de pontos ou de regiões delimitadas oferece maior resolução do QUPC, conseqüentemente proporciona resultados mais precisos. Contudo, o tempo computacional bem como a complexidade na implementação serão maiores, o que é confirmado por [Da Costa, Da Silva Filho, Abe, et al., 1999].

Neste estudo utilizar-se-á o quadrado unitário do plano cartesiano de resolução 12^1 definido em [Da Costa, Da Silva Filho, Abe, et al., 1999], por apresentar uma resolução satisfatória para o propósito de recuperação de informação. Convêm ressaltar que a subdivisão é totalmente livre, sendo a determinação da resolução mais adequada uma tarefa do especialista que irá aplicá-la.

A figura 5.1 a seguir apresenta o quadrado unitário do plano cartesiano de resolução 12^1.

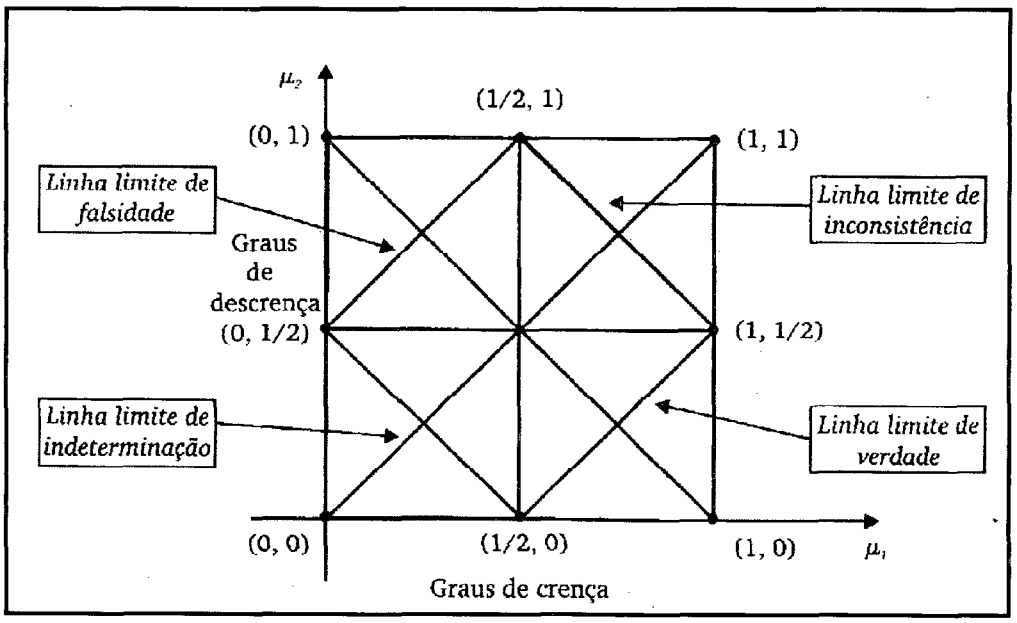


Figura 5.1 – QUPC de resolução 12^1.

No QUPC de resolução 12^1 cada uma das regiões demarcadas pelas linhas recebe uma denominação de acordo com sua proximidade com os pontos extremos do retículo, salienta-se que no retículo representativo da $LPA2_V$ os pontos extremos são: indeterminado, inconsistente, falso e verdadeiro. As regiões relacionadas com os estados ou pontos extremos serão apresentadas a seguir.

Com base nos resultados da $LPA2_V$ representados no retículo, quando o grau de crença e o grau de descrença apresentam valores próximos ou iguais a um (1) resultam num estado inconsistente. Por outro lado se estes apresentarem valores próximos ou igual a zero (0), o estado lógico resultante é indeterminado.

Assim define-se a região totalmente inconsistente (figura 5.2) quando o valor do grau de inconsistência for maior ou igual a 0.5, ou seja, quando o grau de crença e descrença possuírem maiores ou igual a 0.5.

Da mesma forma define-se a região totalmente indeterminada (figura 5.3) quando os valores do grau de crença e descrença resultam num grau de indeterminação igual ou menor a 0.5.

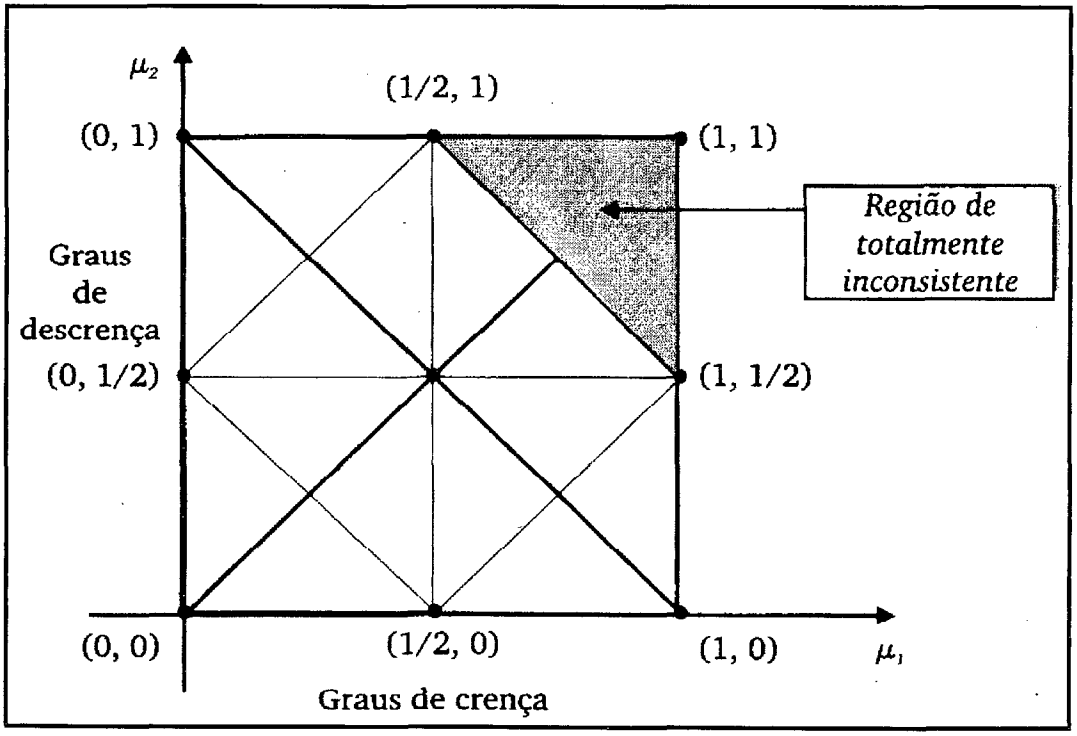


Figura 5.2- QUPC destacando a região totalmente indeterminada

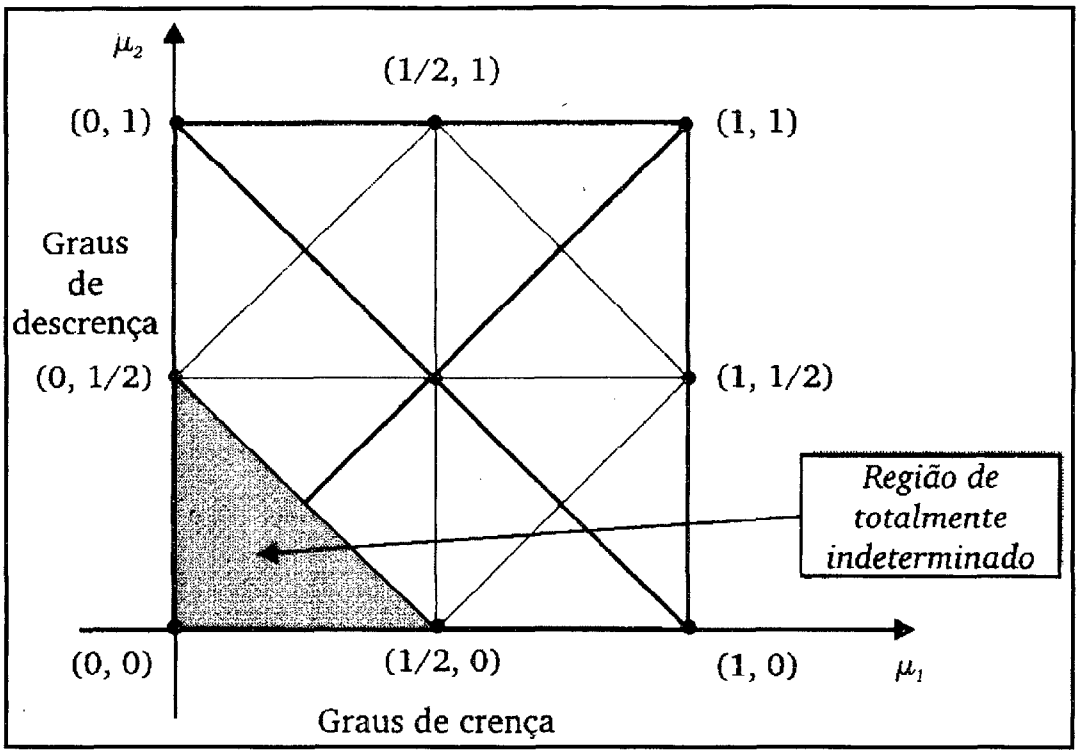


Figura 5.3- QUPC destacando a região Totalmente Indeterminada

De acordo com os resultados da $LPA2_V$ representados no retículo, o resultado será falso se o grau de crença for baixo e o grau de descrença for alto.

No retículo é considerado um estado falso quando os valores dos graus de crença e descrença resultam num grau de falsidade de valor igual ou menor a 0.5. A figura 5.4 apresenta a região de totalmente falso.

Ainda considerando os resultados da $LPA2_V$ representados no reticulado, o resultado será verdadeiro quando o grau de crença for alto, maior ou igual a 0.5, e o grau de descrença for baixo, ou seja, igual ou menor a 0.5. A região considerada totalmente verdadeira é ilustrada na figura 5.5.

Além das regiões que representam os estados extremos (verdadeiro, falso, inconsistente e indeterminado) do retículo, serão apresentadas as regiões que são definidas por sua proximidade com estas regiões.

Próximo a região totalmente falsa e à direita da linha totalmente indefinida localiza-se a região denominada quase falsa, tendendo à inconsistência, representada por: $Qf \rightarrow T$ e sua descrição é: Se $0 \geq G_f \geq -0.5$ e $0.5 < G_{it} \geq 0$ e $\mu_2 > 0.5$ e $0.25 \leq \mu_1 < 0.5$.

A figura 5.6 apresenta a região denominada quase falso, tendendo ao

inconsistente.

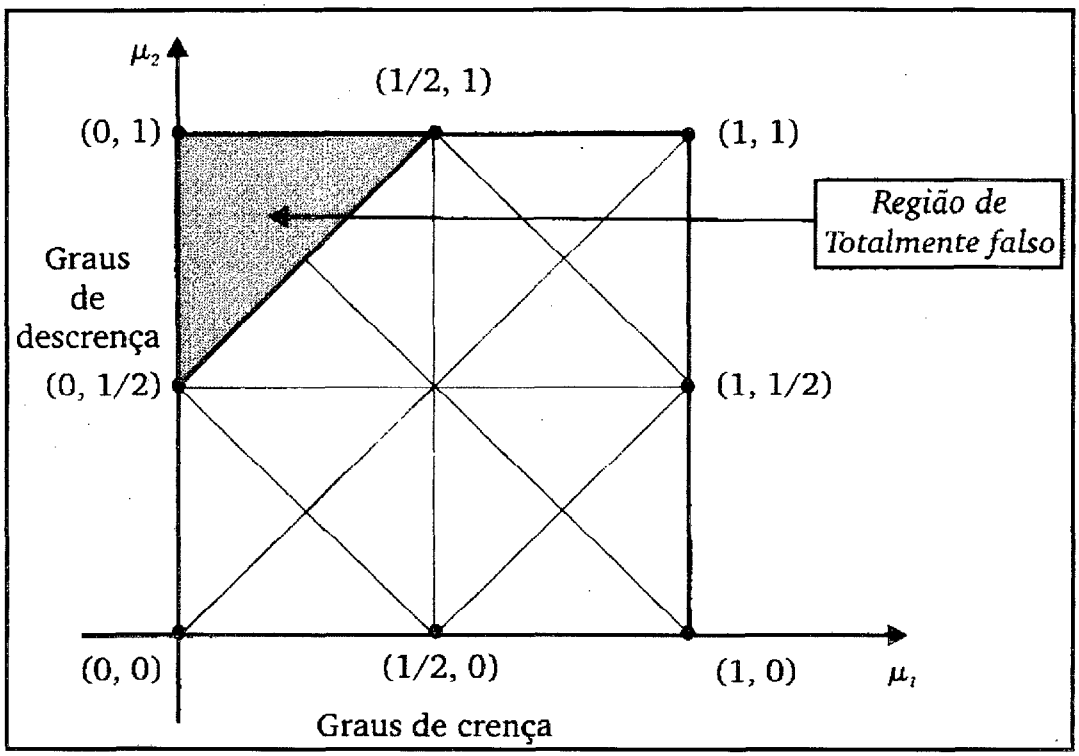


Figura 5.4- QUPC destacando a região Totalmente Falso

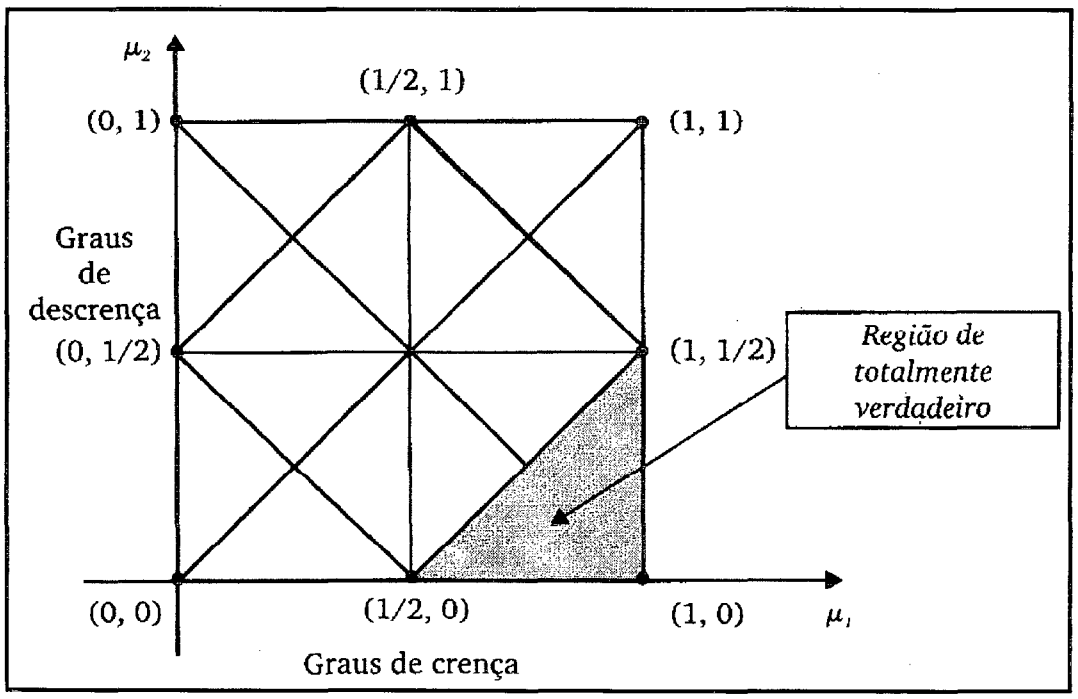


Figura 5.5- QUPC destacando a região totalmente verdadeiro

Já a região localizada próxima à totalmente falsa e à esquerda da linha totalmente indefinida é denominada região quase falsa, tendendo à indeterminação, e é simbolizada

por $Q_f \rightarrow \ \perp$, e sua descrição é: Se $0 \geq G_f \geq -0.5$ e $0.5 < G_{id} \leq 0$ e $0.25 \geq \mu_2 > 0.5$ e $\mu_1 < 0.5$.

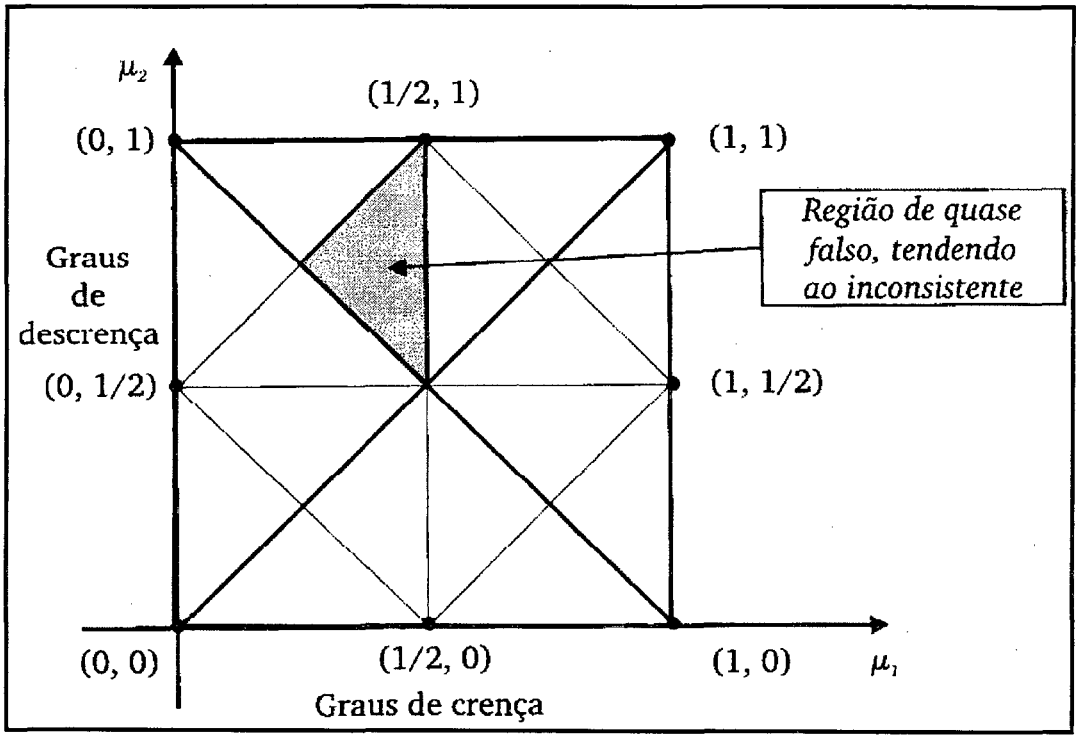


Figura 5.6- QUPC destacando a região quase falso, tendendo ao inconsistente

A figura 5.7 ilustra a região de quase falso, tendendo ao indeterminado.

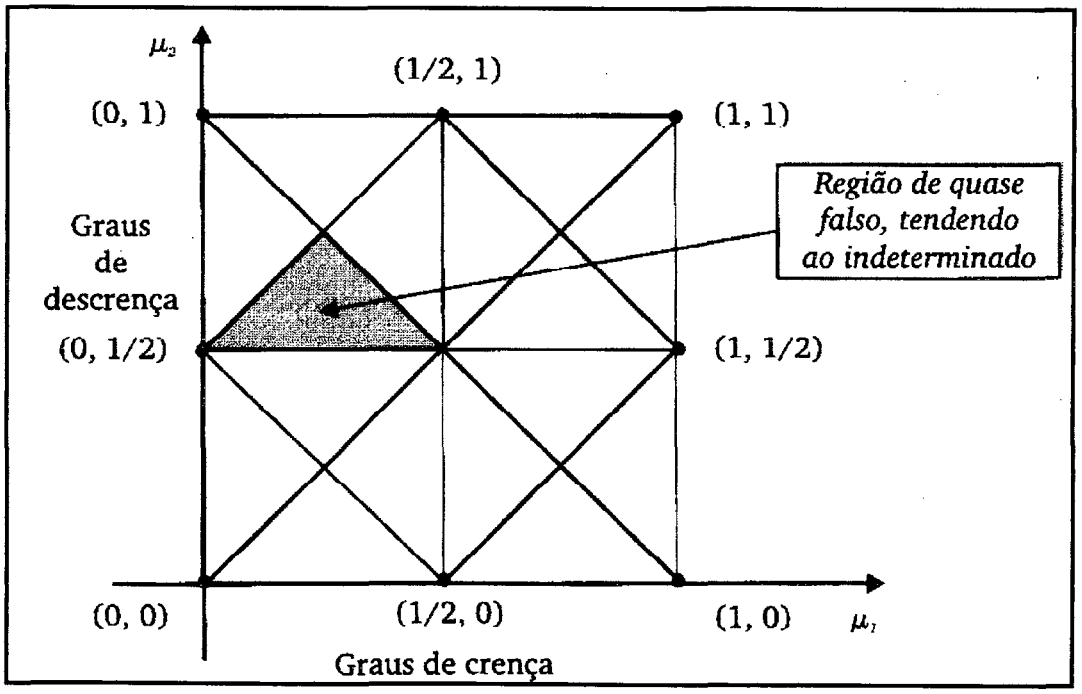


Figura 5.7- QUPC destacando a região de quase falso, tendendo ao indeterminado

A região localizada próxima à totalmente verdadeira e à esquerda da linha totalmente indefinida é denominada região quase verdadeira, tendendo à indeterminação, cuja simbologia é: $Q_v \rightarrow \perp$, e é descrita por: Se $0 \geq G_v > 0.5$ e $-0.5 > G_{id} \leq 0$ e $0.75 \geq \mu_1 > 0.5$ e $\mu_2 < 0.5$.

A região de quase verdadeiro tendendo ao indeterminado é representada pela figura 5.8.

Região quase verdadeira tendendo à inconsistência é aquela localizada próxima à totalmente verdadeira e à direita da linha totalmente indefinida, que é ilustrada na figura 5.9, simbolizada por $Q_v \rightarrow T$ e descrita como segue: Se $0 \geq G_{it} > 0.5$ e $0.5 < G_v \geq 0$ e $\mu_1 > 0.5$ e $0.25 \leq \mu_2 < 0.5$.

Já a região localizada acima da linha totalmente indefinida e próxima à totalmente inconsistente é denominada de região inconsistente tendendo à falsa, cuja simbologia é: $T \rightarrow f$ e é descrito desta maneira: Se $0 \geq G_f > -0.5$ e $0.5 < G_{it} \geq 0$ e $\mu_2 > 0.5$ e $0.75 \leq \mu_1 < 0.5$.

A região inconsistente, tendendo ao falso é apresentada na figura 5.10.

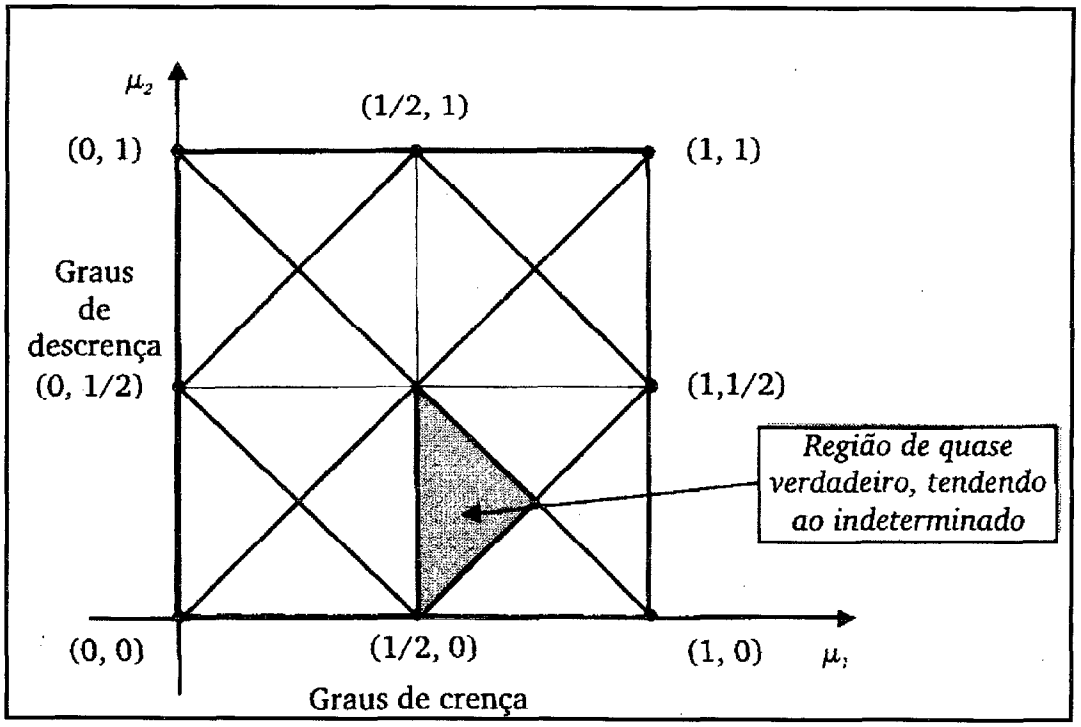


Figura 5.8- QUPC destacando a região de quase verdadeiro, tendendo ao indeterminado

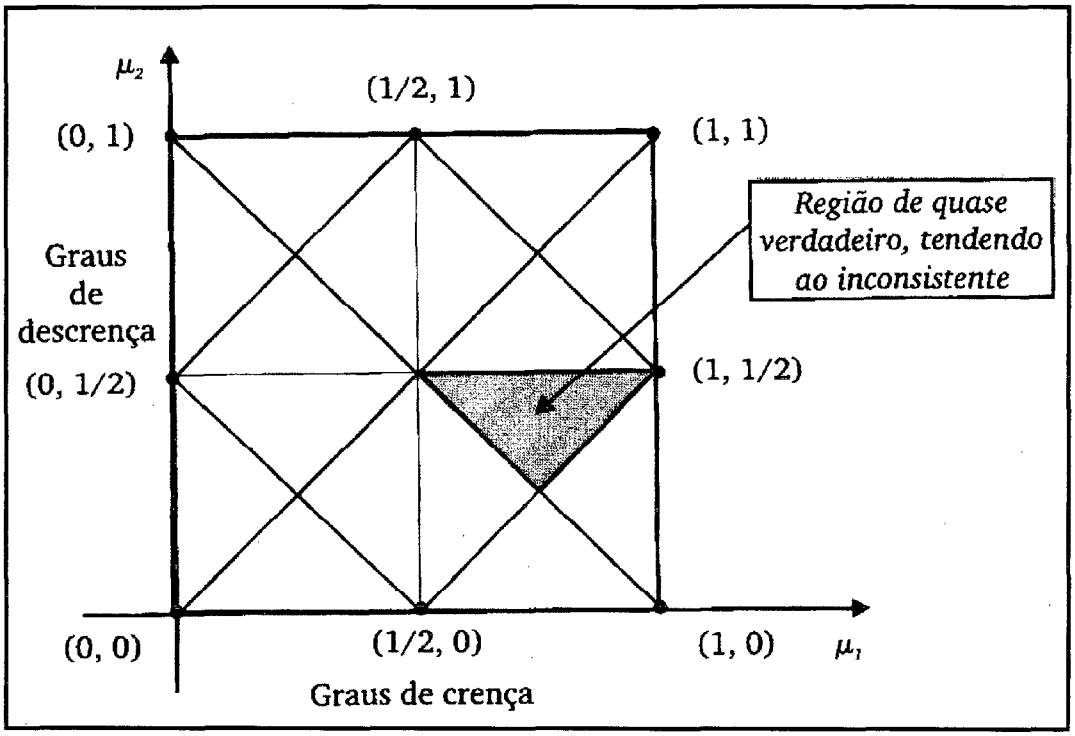


Figura 5.9- QUPC destacando a região quase verdadeiro, tendendo ao inconsistente

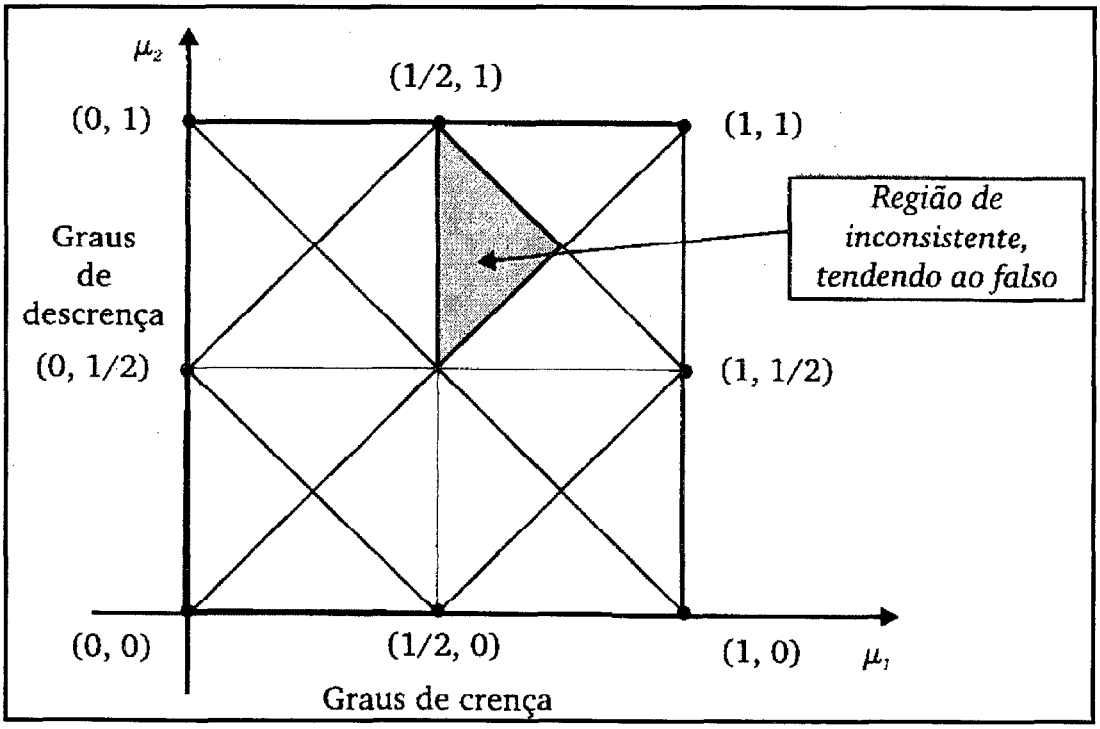


Figura 5.10- QUPC destacando a região de inconsistente, tendendo ao falso

A região denominada de região inconsistente, tendendo à verdadeira é à região abaixo da linha totalmente indefinida e próxima à região inconsistente.

A simbologia do estado inconsistente, tendendo ao verdadeiro é: $T\rightarrow$ v e sua descrição é definida da seguinte forma. Se $0 \geq G_{it} > 0.5$ e $-0.5 < G_v \leq 0$ e $0.75 \geq \mu_2 > 0.5$ e $\mu_1 > 0.5$, assim à saída é inconsistente, tendendo à verdadeira. Esta região, de inconsistente, tendendo ao verdadeiro é apresentada na figura 5.11.

Localizada próxima à região totalmente indeterminada e abaixo da linha totalmente indefinida e acima da linha totalmente definida está localizada a região denominada de região indeterminada tendendo à falsa (figura 5.12), que é simbolizada por: $\perp \rightarrow f$. A descrição da região indeterminada tendendo à falsa é representada da seguinte forma: Se $0 \geq G_f > 0.5$ e $-0.5 > G_{id} \leq 0$ e $0.5 \geq \mu_2 > 0.25$ e $\mu_1 < 0.5$.

E a região que denominada de região indeterminada, tendendo à verdadeira é a região localizada abaixo da linha totalmente indefinida, e abaixo da linha totalmente definida e próxima à totalmente indeterminada (figura 5.13).

A representação simbólica da região de indeterminada, tendendo ao verdadeiro é: $\perp \rightarrow$ v. A descrição da região indeterminada tendendo à verdadeiro é descrita como segue: Se $0 \geq G_{id} > -0.5$ e $0.5 > G_v \geq 0$ e $\mu_2 < 0.5$ e $0.25 \leq \mu_1 < 0.5$, desta forma a saída é indeterminada, tendendo à verdadeira.

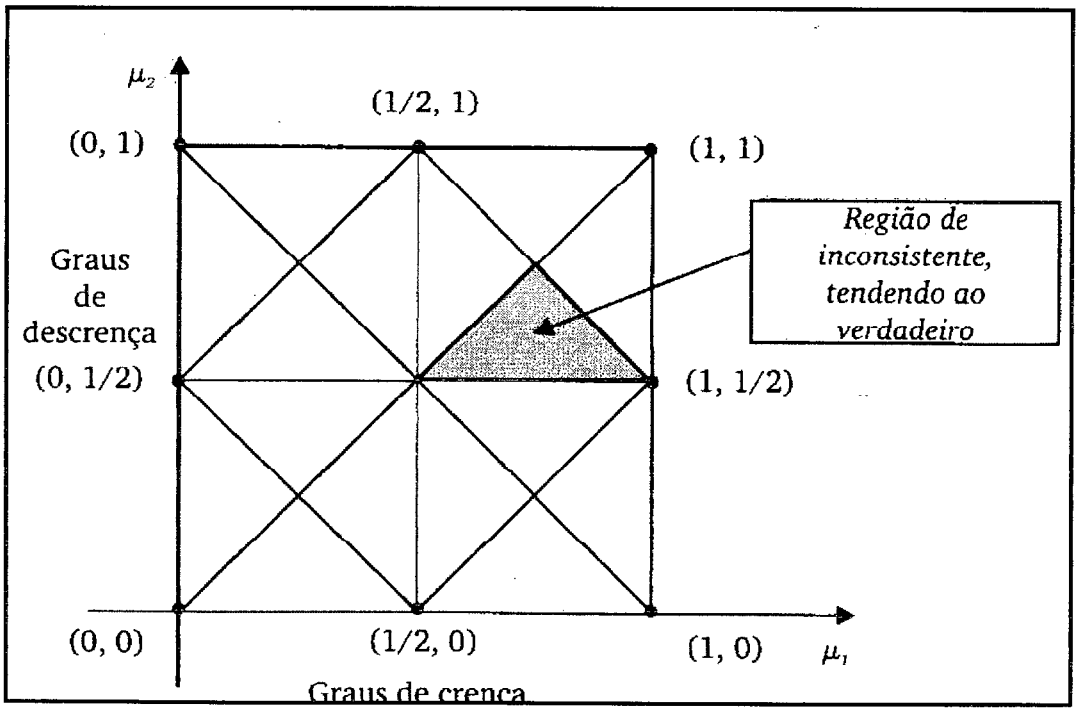


Figura 5.11- QUPC destacando a região de Inconsistente, tendendo ao verdadeiro

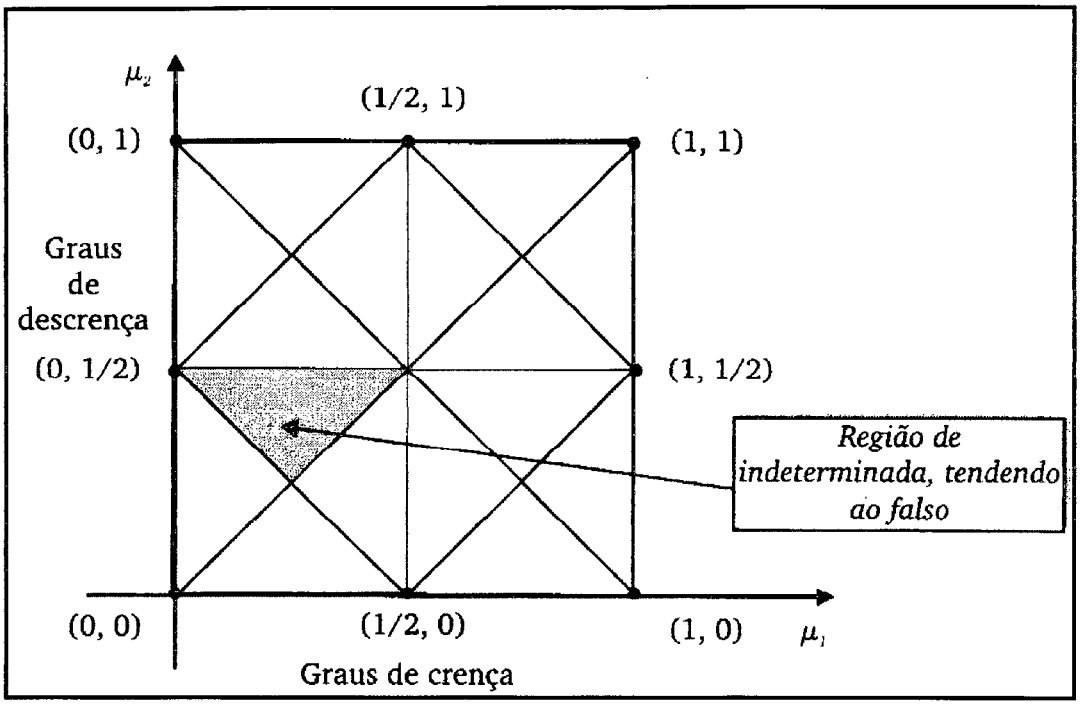


Figura 5.12- QUPC destacando a região de indeterminado , tendendo ao falso

A figura abaixo representa a região de indeterminada, tendendo à verdadeira.

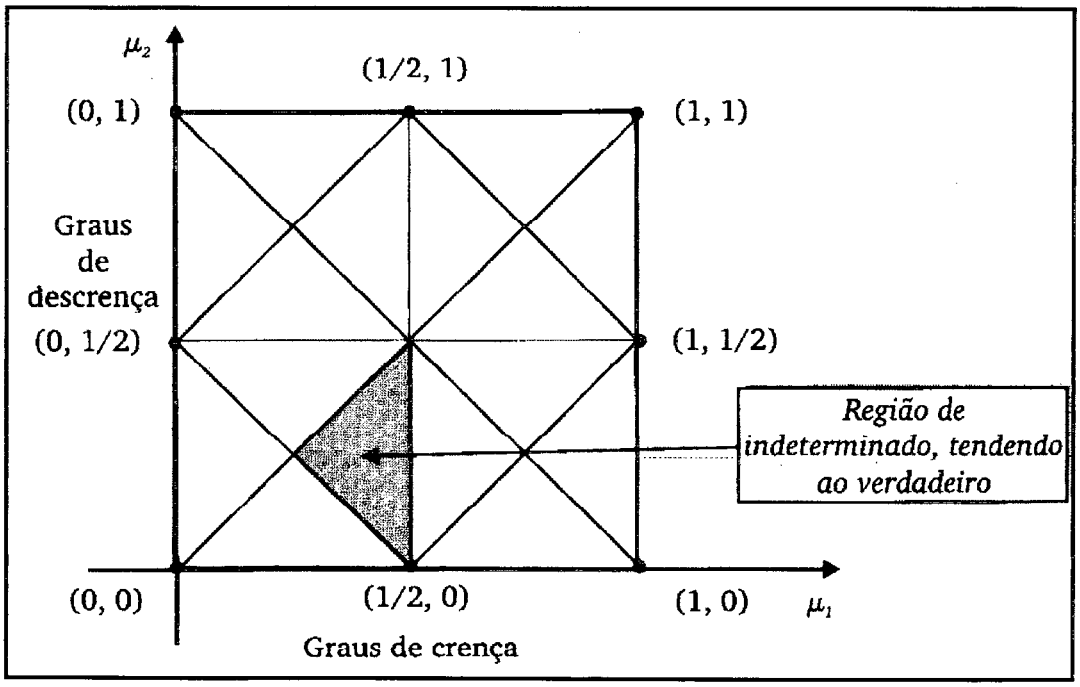


Figura 5.13- QUPC destacando a região de indeterminado, tendendo ao verdadeiro

Essa configuração resultou em 12 regiões do quadrado unitário do plano cartesiano, caracterizando assim o *quadrado unitário do plano cartesiano de resolução 12*.

Conforme [Da Costa, Da Silva Filho, Abe, et al., 1999], na representação das regiões delimitadas no QUPC de resolução 12, são consideradas as seguintes condições:

a) Um grau de inconsistência maior ou igual a 0.5 resulta no estado extremo totalmente inconsistente (Se $G_{it} \geq 0.5 \rightarrow T$); e

b) Um grau de indefinição de valor, em módulo, maior ou igual 0.5 resulta no estado extremo de totalmente indeterminado (Se $G_{id} \geq |0.5| \rightarrow \perp$);e

c) Um grau de verdade maior ou igual a 0.5 resulta no estado extremo totalmente verdadeiro, isto é: (Se $Gv \geq 0.5 \rightarrow V$); e

d) Um grau de falsidade de valor maior ou igual a 0.5, em módulo, resulta no estado extremo totalmente falso. (Se $G_f \geq |0.5| \rightarrow F$).

Com base nestas considerações, os valores em módulos maiores ou iguais a 0.5 já possuem estados lógicos definidos, que são os estados extremos, enquanto que os valores inferiores a 0.5 são inter-relacionados e irão definir regiões que resultaram nos estados lógicos não extremos.

[Da Costa, Da Silva Filho, Abe, et al., 1999] referem que para melhorar a descrição são definidas quatro novas variáveis, que são:

a) Valor superior de controle de certeza (V_{SCC}) que é o valor que irá limitar o grau de certeza próximo ao verdadeiro; e

b) Valor inferior de controle de certeza (V_{ICC}) que representa o valor que vai limitar o grau de certeza próximo ao falso; e

c) Valor superior de controle de contradição (V_{SCCT}) que é o valor que limita o grau de contradição próximo ao estado inconsistente; e

d) E o valor que limita o grau de contradição próximo ao indeterminado é denominado valor inferior de controle de contradição (V_{ICCT}).

De acordo com [Da Costa, Da Silva Filho, Abe, et al., 1999], os valores dos graus de certeza (G_C) e contradição (G_{CT}) representados no reticulado da $LPA2_V$ na configuração do QUPC de resolução 12, ficam com os valores ajustados da seguinte maneira:

$V_{SCC} = 0.5$ $V_{ICC} = -0.5$ $V_{SCCT} = 0.5$ $V_{ICCT} = -0.5$

Todos os estados lógicos do reticulado com os valores de controle considerados nessa configuração são representados na figura 5.14.

O reticulado de forma simplificada para facilitar a visualização, das

representações dos estados extremos e não extremos pode ser feito de forma simbólica, como pode ser observado na figura 5.15.

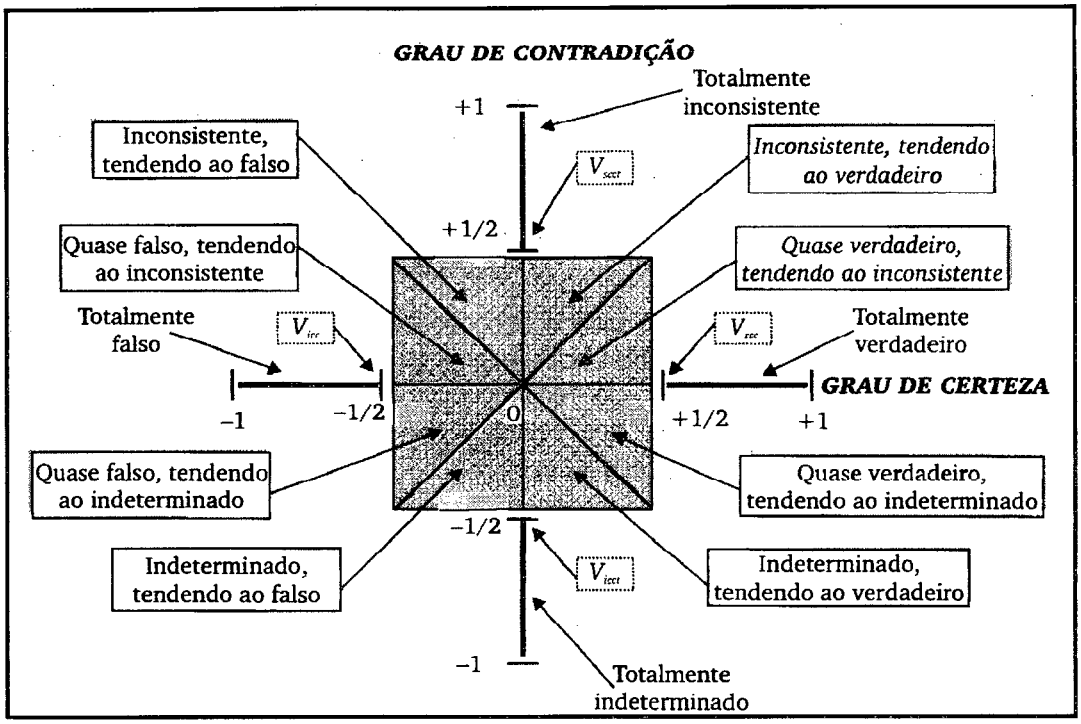


Figura 5.14- Representação dos estados extremos e não extremos com $V_{SCC}=V_{SCCT}$ = 1/2 e $V_{ICC} = V_{ICCT}$ = -1/2

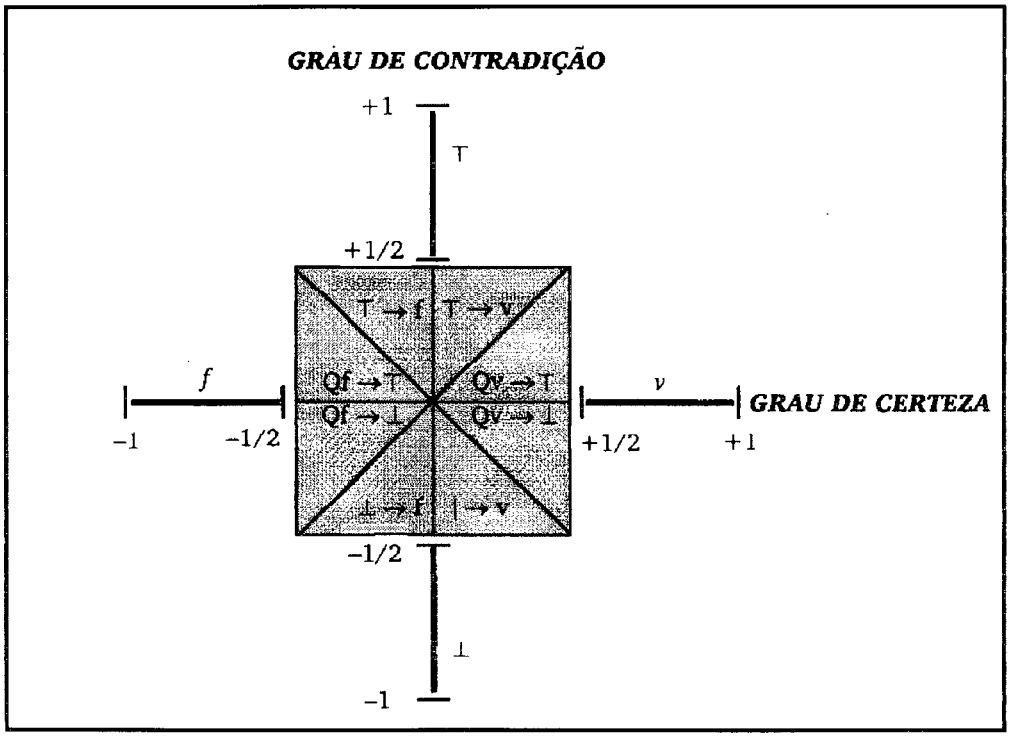


Figura 5.15- Representação simbólica dos estados extremos e não extremos com $V_{SCC}=V_{SCCT}$ = 1/2 e $V_{ICC} = V_{ICCT}$ = -1/2

5.5 VOCABULARIO CONTROLADO E LINGUAGEM NATURAL

Devido ao crescimento do volume de informações armazenadas e das formas de utilização das mesmas em sistemas de recuperação de informação, inúmeras pesquisas foram e estão sendo realizadas pelos mais diversos autores em recuperação. Algumas destas pesquisas dedicam-se a utilização de linguagens: natural e controlada, aplicadas a recuperação e indexação de informação.

Em uma base de dados os atributos (titulo e resumo) referem-se aos termos em linguagem natural, por outro lado os termos de indexação dizem respeito aos termos em linguagem controlada ou vocabulário controlado. Linguagem (vocabulário) controlada nada mais é do que o conjunto (limitado) de termos autorizados para uso na busca e indexação de documentos da coleção.

Em [Raitt, 1980] pode ser encontrado uma revisão de algumas linguagens controladas, dentre elas: NASA Thesaurus, Thesaurus of Engineering and Scientific Terms, Thesaurus of Metallurgical Terms, Subject Headings for Engineering, INIS Thesaurus e INSPEC Thesaurus.

Vários estudos comparativos realizados dos métodos de busca para recuperação de informação envolvendo linguagem natural e linguagem (vocabulário) controlada tais como: [Carrow e Nungent, 1977], [Henzler, 1978] e [Calkins, 1980] apontam para praticamente o mesmo resultado. Esse resultado é que os dois métodos devem ser utilizados como complemento um do outro. Conseqüentemente, a estratégia de busca que utiliza os dois métodos concomitantemente possuirá a melhor performance.

5.6 ABORDAGEM PARACONSISTENTE

5.6.1 Introdução

Nas últimas décadas vem ocorrendo um grande crescimento na quantidade de informações disponíveis em meio digital, diante deste torna-se necessário o desenvolvimento de mecanismos eficientes de busca por informações que supram necessidades específicas.

Os avanços tecnológicos tais como computadores rápidos, dispositivos de

armazenamento com grande capacidade e de alta performance e redes de computadores de alta velocidade não são suficientes. Assim conforme [Baeza-Yates, 1996] e [Ribeiro-Neto e Yates-Baeza, 1999], faz-se necessário à existência de técnicas de armazenamento, acesso, pesquisa e manipulação de dados mais eficientes.

De acordo com [Frakes e Baeza-Yates, 1992] muitas ferramentas para recuperação de informação foram e estão sendo desenvolvidas, dentre as quais podem-se destacar os índices, que são uma coleção de termos com ponteiros para lugares onde informações relacionadas a estes podem ser encontradas.

[Frakes e Baeza-Yates, 1992] referem que índices são extremamente simples e eficientes para uso em coleções de dados pequenas e médias.

5.6.2 Ajustes no QUPC para o modelo de recuperação paraconsistente

Conforme referido anteriormente o diagrama da $LPA2_V$ é construído a partir dos valores dos graus de certeza (verdade e falsidade) e de contradição (inconsistência e indeterminação), cuja análise destes é realizada através do quadrado unitário do plano cartesiano (QUPC), que utiliza forma inicial, da configuração e dos valores considerados pelo algoritmo "para-analisador" descrito na seção 3.5.

De acordo com [Da Costa, Da Silva Filho, Abe, et al., 1999], através da análise do algoritmo pode-se observar que a alteração de qualquer um dos quatro valores de controle de limites (C_1,C_2, C_3 ou C_4) resulta na alteração das características do estado lógico resultante como saída do sistema.

Essas alterações podem ser mais bem visualizadas através do QUPC, onde é fácil verificar que ao serem realizadas alterações dos valores de controle limites, ocorrerá uma mudança nas dimensões das relacionadas com os estados lógicos resultantes de saída.

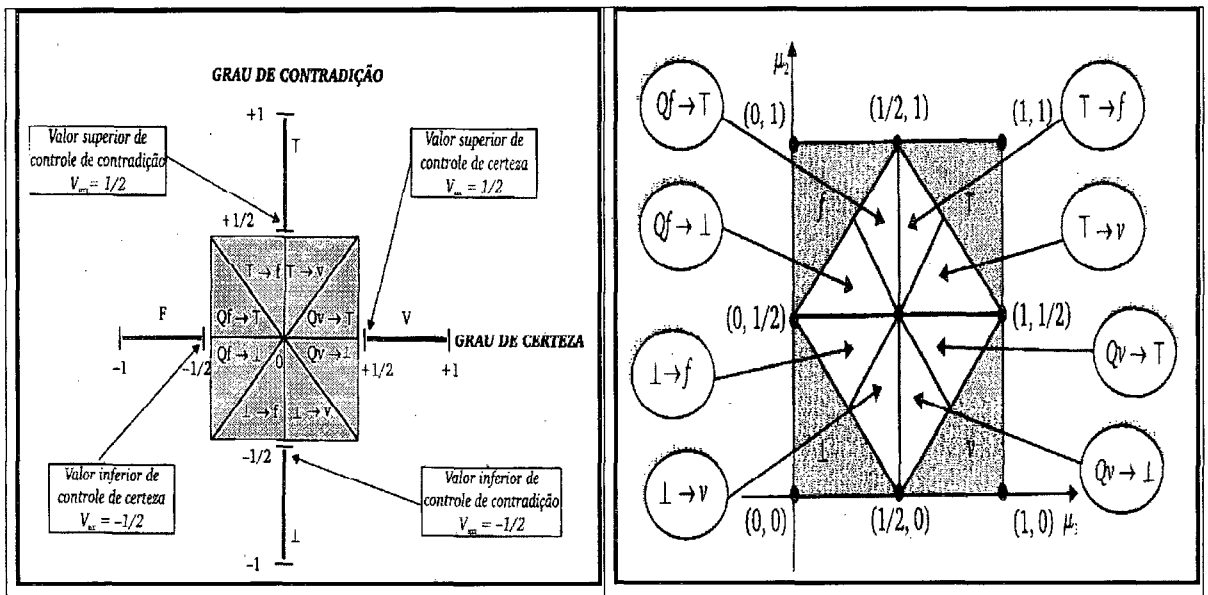
Para melhor visualização destas alterações, a seguir será apresentado o reticulado da LPA2V representado pelo gráfico dos graus de certeza e contradição e as modificações realizadas nas regiões delimitadas no QUPC para alguns valores de controle limite diferentes.

As figuras do reticulado construído com os valores do grau de certeza e contradição serão apresentadas juntamente com as figuras das regiões delimitadas no QUPC para melhor visualização dos efeitos proporcionados pelas alterações dos valores de controle extremos.

O reticulado com os valores dos graus de certeza e contradição representados no QUPC da figura 5.16 apresentam os valores de controle iguais a ± ½, essa configuração é a utilizada pelo QUPC de resolução 12.

Uma outra resolução, mais sensível é demonstrada na figura 5.17 onde os valores de controle: $V_{SCC} = V_{SCCT} = 1/4$ e $V_{ICC} = V_{ICCT} = -1/4$. Tal sistema é mais sensível a situações que vão resultar em estados extremos.

O exemplo da figura 5.18 os valores de controle são ajustados em ± 3/4, ($V_{SCC} = V_{SCCT} = 3/4$ e $V_{ICC} = V_{ICCT} = -3/4$). Ao contrário do exemplo anterior, figura 5.17, os sistemas com os valores de controle 3/4 são mais sensíveis a situações que resultam nos estados não extremos.

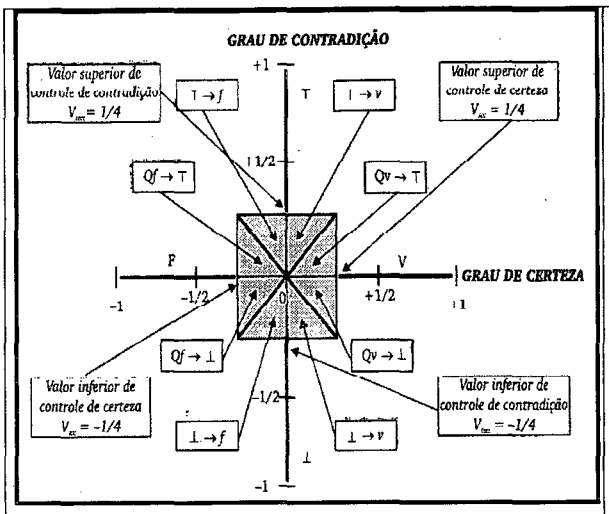


a) Representação no gráfico dos graus de certeza e de contradição com os valores de controle de limite: $V_{SCC} = S_{SCCT} = 1/2$ e $V_{ICI} = V_{ICCT} = -1/2$

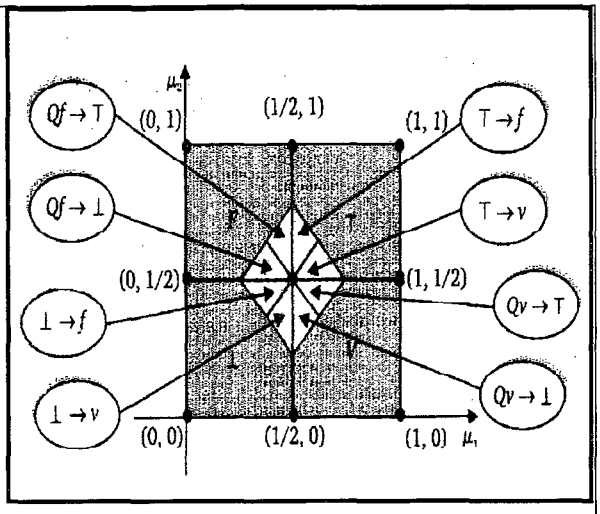
b) Representação no QUPC das regiões dos estados extremos e não-extremos com os valores de controle $V_{SCC} = S_{SCCT} = 1/2$ e $V_{ICI} = V_{ICCT} = -1/2$

Figura 5.16 – Representação em graus e no QUPC das regiões dos estados extremos e não extremos com os valores de controle ajustados em ± 1/2

Até agora foram apresentadas situações onde os valores de controle eram iguais nas figuras 5.19 e 5.20 será apresentados situações nas quais os valores certeza e contradição são diferentes.

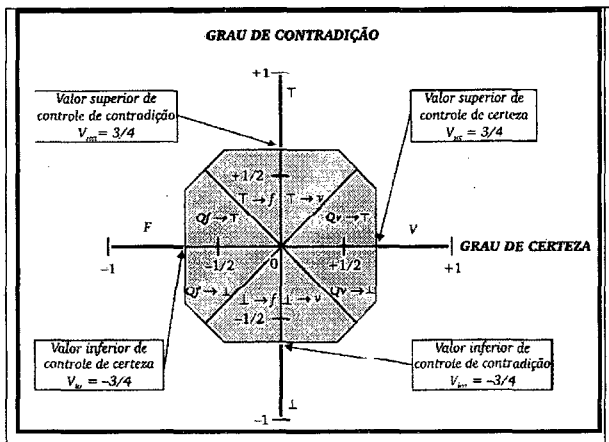


a) Representação no gráfico dos graus de certeza e de contradição com os valores de controle de limite: $V_{SCC} = S_{SCCT} = 1/4$ e $V_{ICI} = V_{ICCT} = -1/4$

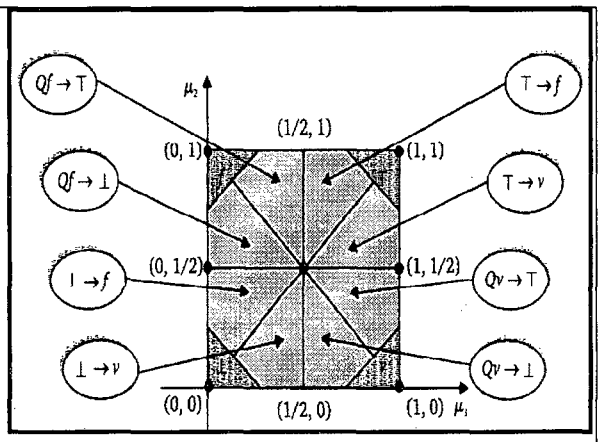


b) Representação no QUPC das regiões dos estados extremos e não-extremos com os valores de controle $V_{SCC} = S_{SCCT} = 1/4$ e $V_{ICI} = V_{ICCT} = -1/4$

Figura 5.17 – Representação em graus e no QUPC das regiões dos estados extremos e não extremos com os valores de controle ajustados em ± 1/4



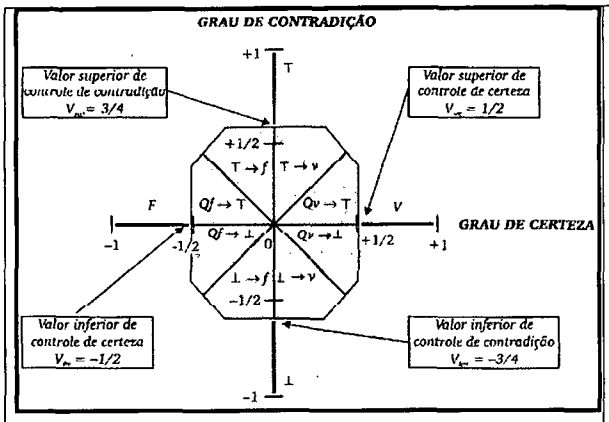
a) Representação no gráfico dos graus de certeza e de contradição com os valores de controle de limite: $V_{SCC} = S_{SCCT} = 3/4$ e $V_{ICI} = V_{ICCT} = -3/4$



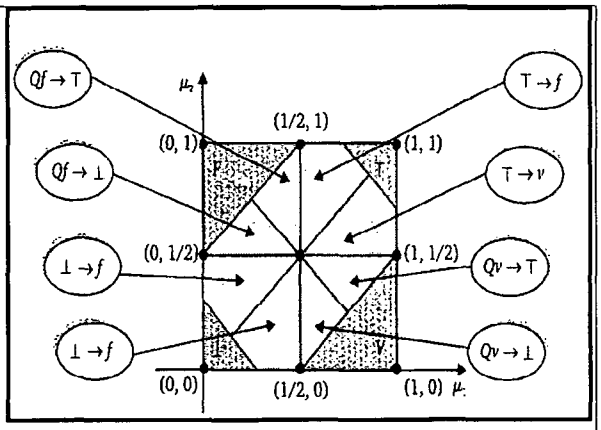
b) Representação no QUPC das regiões dos estados extremos e não-extremos com os valores de controle $V_{SCC} = S_{SCCT} = 3/4$ e $V_{ICI} = V_{ICCT} = -3/4$

Figura 5.18– Representação em graus e no QUPC das regiões dos estados extremos e não extremos com os valores de controle ajustados em ± 3/4

Para obter-se uma maior sensibilidade a situações que resultam em estados extremos de certeza do que estados extremos de contradição os valores de controle são ajustados da seguinte forma: $V_{SCC} = V_{ICC} = \pm 1/2$ e $V_{SCCT} = V_{ICCT} = \pm 3/4$, como pode ser observado na figura 5.19. Conseqüentemente, tal configuração é menos sensível a situações que resultam em estados não extremos de certeza do que estados não extremos de contradição.



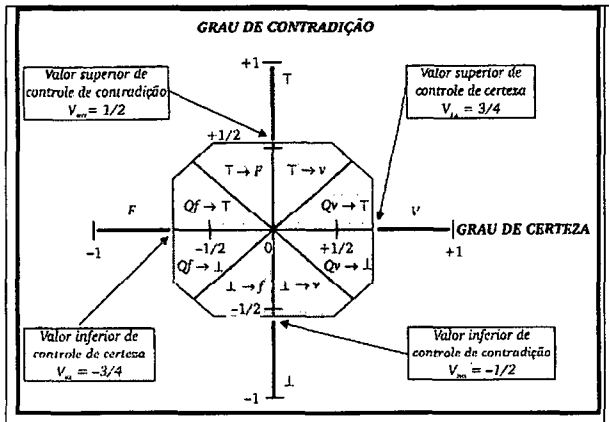
a) Representação no gráfico dos graus de certeza e de contradição com os valores de controle de limite: $V_{SCC} = S_{SCCT} = \pm ½$ e $V_{ICI} = V_{ICCT} = \pm 3/4$



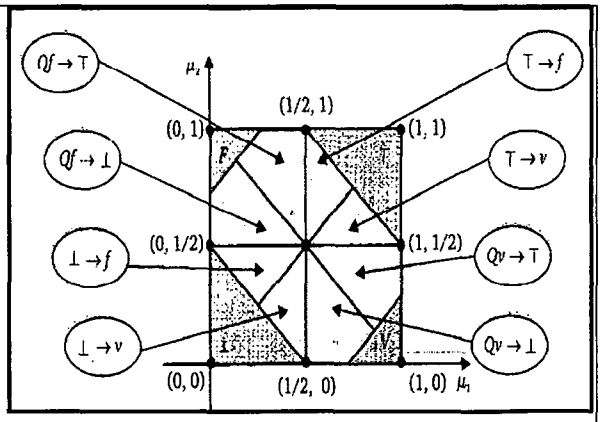
b) Representação no QUPC das regiões dos estados extremos e não-extremos com os valores de controle $V_{SCC} = S_{SCCT} = \pm 1/2$ e $V_{ICI} = V_{ICCT} = \pm 3/4$

Figura 5.19 – Representação em graus e no QUPC das regiões dos estados extremos e não extremos com os valores de controle: $V_{SCC} = V_{ICC} = \pm 1/2$ e $V_{SCCT} = V_{ICCT} = \pm 3/4$

Ajustando os valores de controle com: $V_{SCC} = V_{ICC} = \pm 3/4$ e $V_{SCCT} = V_{ICCT} = \pm 1/2$ aumenta-se à sensibilidade a situações que resultam em estados extremos de contradição é ainda menos sensível a situações que resultam em estados não extremos de contradição do que de certeza, como é ilustrado na figura 5.20.



a) Representação no gráfico dos graus de certeza e de contradição com os valores de controle de limite: $V_{SCC} = S_{SCCT} = \pm 3/4$ e $V_{ICI} = V_{ICCT} = \pm ½$

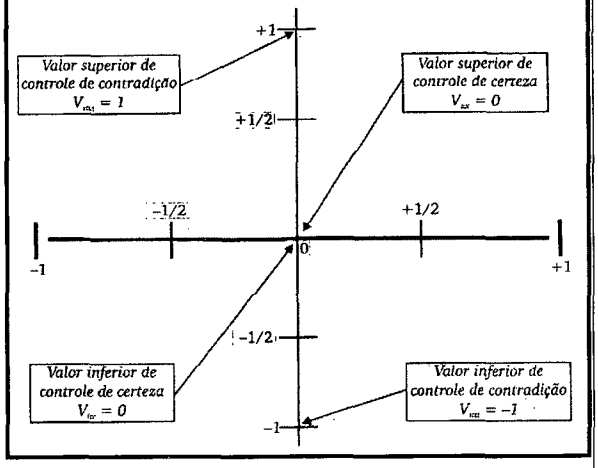


b) Representação no QUPC das regiões dos estados extremos e não-extremos com os valores de controle $V_{SCC} = S_{SCCT} = \pm 3/4$ e $V_{ICI} = V_{ICCT} = \pm 1/2$

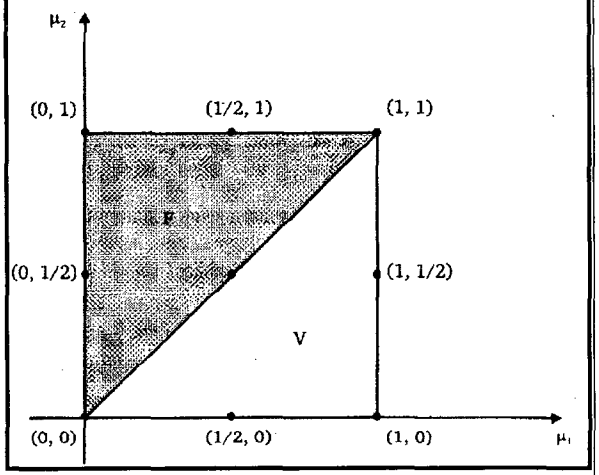
Figura 5.20 – Representação em graus e no QUPC das regiões dos estados extremos e não extremos com os valores de controle: $V_{SCC} = V_{ICC} = \pm 3/4$ e $V_{SCCT} = V_{ICCT} = \pm 1/2$

Em uma situação em que os graus de contradição $V_{SCCT} = 1$ e $V_{ICCT} = -1$ e os valores de certeza são ajustados com zero ($V_{SCC} = V_{ICC} = 0$) observa-se que as regiões que delimitam os estados não extremos desaparecem, como pode ser observado na figura 5.21, assim as regiões dos estados extremos de contradição não existem mais, conseqüentemente os estados extremos de inconsistência e de indeterminação deixam de

ser considerados.



a) Representação no gráfico dos graus de certeza e de contradição com os valores de controle de limite: $V_{SCC} = S_{SCCT} = 0$ e $V_{ICI} = V_{ICCT} = \pm 1$



b) Representação no QUPC das regiões dos estados extremos e não-extremos com os valores de controle $V_{SCC} = S_{SCCT} = 0$ e $V_{ICI} = V_{ICCT} = \pm 1$

Figura 5.21 – Representação em graus e no QUPC das regiões dos estados extremos e não extremos com os valores de controle: $V_{SCC} = V_{ICC} = 0$ e $V_{SCCT} = V_{ICCT} = \pm 1$

De acordo com [Da Costa, Da Silva Filho, Abe, et al., 1999], o sistema de controle ajustado com estes valores é totalmente binário, portanto fica claro que existindo a possibilidade de ajustar o algoritmo desta forma, controle binário, evidencia-se que a Lógica Paraconsistente Anotada engloba a Lógica Clássica.

5.6.3 Modelo paraconsistente para recuperação de informação

Conforme mencionado anteriormente existem várias abordagens adicionais para modelar o processo de recuperação de informação, o que é confirmado por [Ribeiro-Neto e Baeza-Yates, 1999]. Utilizar-se-á neste trabalho a adoção de um *thesaurus*.

A idéia base (fundamental) do modelo aqui proposto é expandir o conjunto de termos indexados na *query* com termos relacionados, obtidos de um *thesaurus*, com o objetivo de aumentar a quantidade de documentos relevantes recuperados para uma *query* do usuário, que não seriam recuperados em a utilização deste recurso.

Thesaurus podem ser construídos a partir de vários processos, ver capitulo quatro. Um destes processos pode ser a partir de uma matriz de similaridade de termos (*term-vs-term similarity matrix)* definida em [Qui and Frei, 1993], que também é utilizada em alguns modelos baseados em lógica *fuzzy* tais como em [Ogawa, Morita e Kobayashi, 1991] no qual é denominada matriz de conexão de palavras-chave.

Nesta matriz, denominada $\vec{c}$ as linhas e colunas são associadas com termos

indexados da coleção de documentos. Na matriz $\vec{c}$ um fator de correlação, entre os termos k_i e k_l, é denominado $c_{i,l}$. Tal fator pode ser definido da seguinte forma:

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

Onde:

n_i, representa o número de documentos que contém o termo k_i;

n_l, o número de documentos que contém o termo k_l;

$n_{i,l}$, o número de documentos que contém ambos os termos (ki, k_l).

Segundo [Ribeiro-Neto e Baeza-Yates, 1999], tal métrica de correlação é bastante comum e tem sido muito utilizada em algoritmos de *clustering*.

O modelo proposto neste trabalho utilizará como grau de crença para um determinado termo K_i o fator de correlação, $c_{i,l}$,do termo K_i.

Convém ressaltar que em função desta pesquisa ser pioneira no estudo da aplicação da lógica paraconsistente no processo de recuperação de informação, bem como da pesquisa bibliográfica não foi possível determinar de forma automática um valor inicial para o grau de descrença do termo k_i, desta forma para fins de análise inicial da relevância do termo, atribuiu-se o valor constante igual a 0.4 para o grau de descrença.

Assim determinados os valores dos graus de crença e descrença do termo k_i em relação ao documento d_j, estes são submetidos para análise paraconsistente através do algoritmo "para-analisador" descrito na seção 3.5. que está configurado para QUPC de resolução 12.

Em decorrência de ter-se fixado o valor do grau de descrença dos termos K haverá conseqüentemente redução dos estados lógicos resultantes.

Após análise inicial dos termos o usuário terá como saída os estados lógicos resultantes bem como os valores dos graus de certeza (verdade e falsidade) e de contradição (indeterminação e inconsistência).

Os termos que possuírem valores dos graus de certeza (verdade e falsidade) e de contradição (indeterminação e inconsistência), situados nas regiões indeterminado tendendo ao verdadeiro, inconsistente tendendo ao verdadeiro, verdadeiro tendendo ao incosnsistente, verdadeiro tendendo ao indeterminado e totalmente verdadeiro (figura

5.22) serão submetidos ao processo de julgamento de relevância (*relevance feedback*) por parte do usuário.

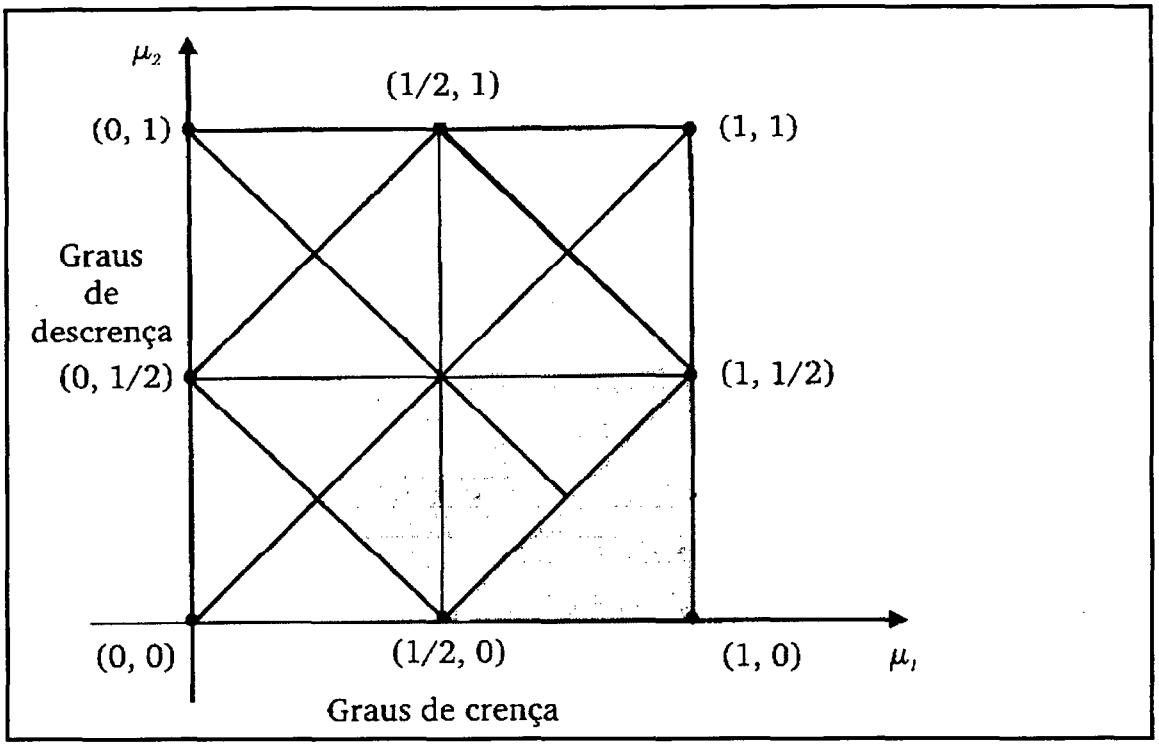


Figura 5.22 – QUPC destacando as regiões submetidas a julgamento do usuário

A partir disso o usuário fará uso da estratégia de *Relevance Feedback*, seção 2.11 através da qual ele irá determinar o grau de descrença dos termos para análise final.

Como resultado obter-se-á a relação dos termos indexado com seus respectivos valores dos graus de certeza e de contradição.

A utilização de *relevance feedback* se dá ao fato de que técnica de construção de *thesaurus* que não fazem uso deste recurso não apresentarem melhora significativa na performance de sistemas de recuperação de informação.

A implementação de um modelo com essas características se comparado ao modelo fuzzy proposto em [Ogawa, Morita e Kobayashi, 1991], que também faz uso da matriz de correlação dos termos para construção de um *thesaurus* apresentará melhora significativa, pois este modelo bem como os demais modelos baseados em lógica *fuzzy* consideram apenas o grau de pertinência do termo K_i em correlação ao documento d_j.

No modelo paraconsistente proposto nesta pesquisa, além de considerar-se o grau de crença que equivale ao grau de pertinência do modelo *fuzzy*, considera-se também o grau de descrença do termo, o que proporciona melhora significativa na qualidade dos

termos que constituem o *thesaurus*.

Isso pode ser confirmado através dos estudos de [Da Silva Filho e Abe, 1999] que referem que na prática um sistema paraconsistente funciona da seguinte forma:

1. Se o grau de contradição for muito elevado, significa que não há certeza ainda quanto à decisão, desta forma é necessário obter-se novas evidências (proposições), *relevance feedback*;
2. Se houver um alto grau de certeza a conclusão pode ser gerada, desde que haja um baixo grau de incerteza.

A tabela 5.1 demonstra a relação entre os estados lógicos resultantes com relação ao grau de crença e descrença da proposição.

INCONSISTENTE	Grau de crença alto
	Grau de descrença alto
VERDADE	Grau de crença alto
	Grau de descrença baixo
FALSO	Grau de crença baixo
	Grau de descrença alto
INDETERMINADO	Grau de crença baixo
	Grau de descrença baixo

Tabela 5.1 – Estado lógico resultante com relação aos graus de crença e descrença

Desta forma fica claro que no caso do modelo *fuzzy* o termo k_i pode ser considerado um bom termo para constituir o *thesaurus*, pois possui alto grau de pertinência, mas não faz nenhuma consideração sobre o seu grau de descrença, e conforme mencionado anteriormente um termo pode ter alto grau de crença e também ter alto grau de descrença sendo considerado um termo não relevante para constituir o *thesaurus*.

Como o objetivo do modelo paraconsistente proposto neste trabalho é melhorar a qualidade dos termos que compõem o *thesaurus*, inicialmente só farão parte do *thesaurus* aqueles termos que obtiverem um alto grau de crença e um baixo grau de descrença, no caso o estado lógico resultante da análise paraconsistente será o estado verdadeiro, conforme por ser observado na figura 5.5.

5.6.4 Arquitetura do modelo paraconsistente

O sistema proposto, de forma inicial será constituído dos seguintes componentes principais: interface, *relevance feedback,* "para-análise", termos, matriz de correlação, graus de crença e descrença, coleção de documentos, conforme pode ser observado na figura 5.23.

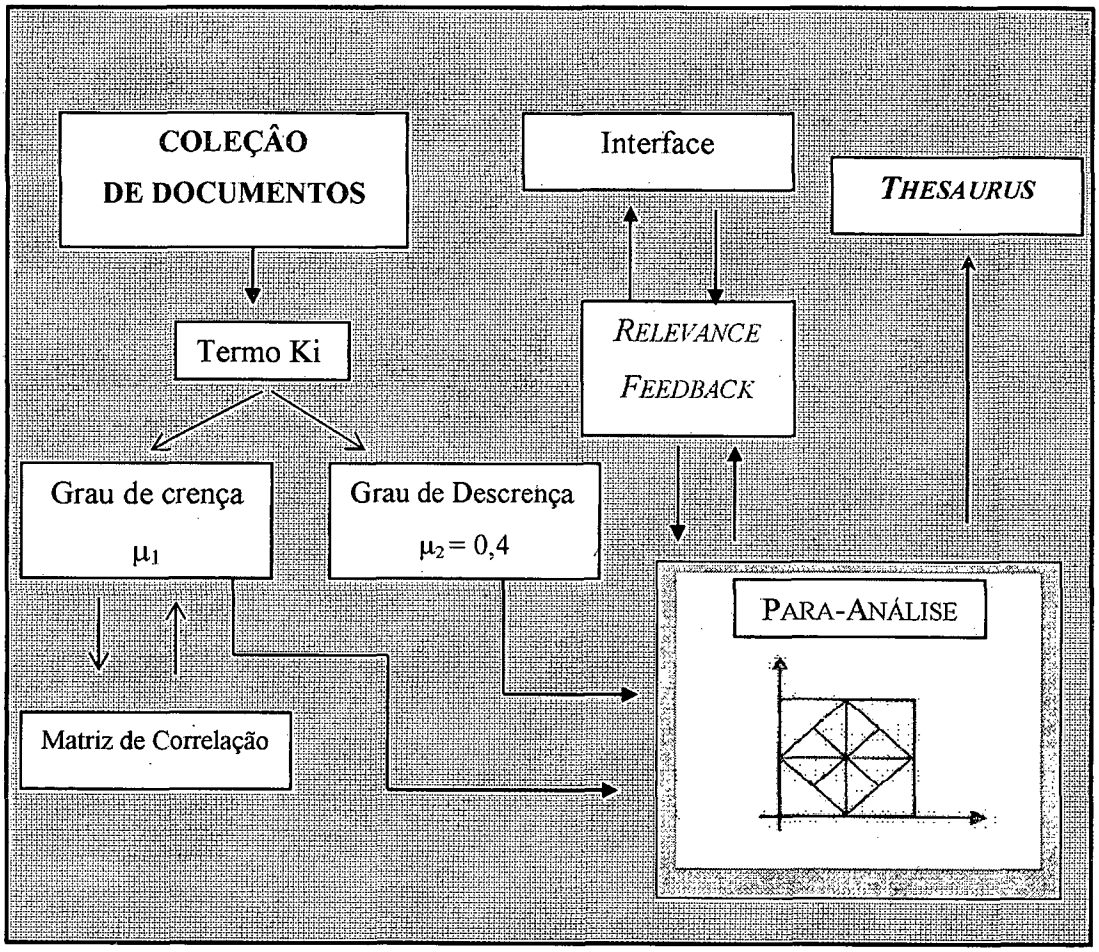


Figura 5.23: Arquitetura modelo paraconsistente

Onde:

Interface: através da qual o usuário do sistema de recuperação de informação fará o julgamento da relevância do termo K_i para a sua *query;*

Termo K_i: cada um dos termos;

Grau de crença: o grau de crença atribuído ao termo K_i, obtido através da matriz de correlação;

Grau de descrença: o grau de descrença atribuído ao termo K_i. Este será

inicializado com o valor de 0,4;

Matriz de correlação: responsável pela geração do grau de relevância do termo K_i;

Para-análise: Através dos graus de crença e descrença do termo K_i calcula o grau de certeza e de contradição do termo que resultará em um estado lógico como saída; algoritmo para-analisador;

Relevance Feedback: retorna ao usuário após pré-análise os graus de crença e descrença do termo K_i, para análise por parte do usuário que irá readequar os pesos destes para sua necessidade de informação (*query).*

Coleção de documentos: coleção de documentos a partir dos quais serão gerados os termos que constituirão o *thesaurus.*

Conforme referido anteriormente, sistema de recuperação de informação (SRI) podem ser avaliados através de consultas que fazem parte de uma coleção de referência, desta forma após a implementação do modelo proposto nesta pesquisa pretende-se submeter tal modelo para avaliação em uma coleção de referência tal como a *Text Retrieval Conference* (TREC).

Além disso, futuramente pretende-se realizar o estudo de um modo de obtenção de forma automática do valor do grau de descrença de um termo, proporcionando desta forma melhora da performance do sistema.

6 CONCLUSÕES

Nesta pesquisa realizou-se um amplo estudo teórico a cerca das várias abordagens para recuperação de informação com o propósito de verificar a possibilidade da utilização da lógica paraconsistente como uma abordagem para recuperação de informação.

Vários modelos foram e têm sido propostos para trabalhar com recuperação de dados em bases textuais. Entre eles estão: o modelo vetorial proposto por Salton apud [Lima, Laender, Ribeiro-Neto, 1998], o modelo hierárquico proposto por [Lima, Laender e Ribeiro-Neto, 1998], o modelo baseado em rede bayesiana proposto por [Silva e Ribeiro-Neto, 1998] e o modelo baseado em filtros proposto por [Mendonça, Silva e Ribeiro-Neto, 1998].

Os sistemas de recuperação de informação de modo geral estão sofrendo cada vez mais com a ineficiência no seu processo de indexação e busca da informação. Isso se deve principalmente porque nenhum tratamento adequado sobre situações que apresentem inconsistências é empregado a tais sistemas.

A razão pela qual optou-se por propor um modelo para recuperação de informação em bases textuais utilizando lógica paraconsistente decorre da facilidade proporcionada por está no tratamento de incertezas, paradoxos, inconsistências e vagacidade (*vagueness)*, uma vez que bases textuais são repletas destas situações, pois em sua maioria fazem uso de linguagem natural.

No presente trabalho realizou-se uma análise teórica preliminar do algoritmo "para-analisador", através da qual obtem-se se o termo é relevante ou não para representar o documento, e se tal termo for considerado relevante deverá fazer parte do *thesaurus*. As tarefas de definir que termos melhor representarão a *query* e os documentos apresentam grande complexidade.

[Ribeiro-Neto e Baeza-Yates, 1999] referem que existem várias abordagens adicionais para modelar o processo de recuperação de informação, neste trabalho como abordagem adicional foi utilizado *thesaurus*.

As principais vantagens do uso do vocabulário controlado (*thesaurus*) na recuperação de informação mencionadas por [Henzler, 1978], [Perez (1982] e [Salton , 1986], são sintetizadas a seguir.

1. Controle total do vocabulário de indexação minimiza os problemas de comunicação entre indexadores e usuários.

2. Com o uso de um *thesaurus* e suas respectivas notas de escopo os indexadores podem assinalar mais corretamente os conceitos dos documentos.

3. As relações hierárquicas e remissivas do vocabulário controlado auxiliam o indexador e o usuário na identificação de conceitos relacionados.

4. Se bem constituído, o vocabulário controlado poderá oferecer alta recuperação e relevância.

5. Redução no tempo de consulta à base de dados, pois a estratégia de busca será melhor elaborada com o uso do *thesaurus*.

Thesaurus podem ser construídos de duas formas: manualmente ou automaticamente.

Thesaurus construídos automaticamente são baseados na co-ocorrência de informação e no julgamento de relevância, e geralmente são utilizados para estimar a probabilidade que o termo do *thesaurus* possui de apresentar similaridade com os termos da *query*.

A idéia central é expandir o conjunto de termos indexados na *query* com termos relacionados, obtidos de um *thesaurus*, com o objetivo de melhorar a quantidade e qualidade dos documentos recuperados para uma dada *query*.

Alguns modelos baseados em lógica *fuzzy* foram propostos para construção de *thesaurus* automaticamente, nos quais é encontrado o grau de pertinência do termo com relação aos documentos da coleção.

Percebeu-se se através da comparação do modelo proposto nesta pesquisa com o modelo *fuzzy* proposto por [Ogawa, Morita e Kobayashi, 1991], que o modelo *fuzzy* apresenta desvantagens uma vez que este só considera o grau de pertinência do termo. Desta forma termos indexados contraditórios podem ser considerados relevantes, pois não há neste modelo nenhuma forma de controlar situações contraditórias ou inconsistentes.

O modelo apresentado nesta pesquisa considera tanto o grau de crença quanto o grau de descrença do termo, assim proporcionando tratamento adequado para estas situações.

O presente estudo revelou que é possível a utilização da lógica paraconsistente

como abordagem para recuperação de informação em bases textuais, através do emprego desta na construção de *thesaurus*.

Os resultados obtidos se constituem assim, numa contribuição científica original para a área, uma vez que não existe nenhum estudo específico e detalhado para avaliação de técnicas de lógica paraconsistente aplicada à recuperação de informação. Portanto, há necessidade de definição para esta área não só de conceitos, mas também de arquiteturas, processos, técnicas e ferramentas.

Por fim, pode-se afirmar que a área de recuperação de informação baseada em lógica paraconsistente precisa associar-se a outras áreas da Ciência da Computação, como processamento de linguagem natural, inteligência artificial, banco de dados entre outras, mas que também precisará conhecer técnicas advindas de outras áreas tais como Lingüística (para construção do *thesaurus*).

7 REFERÊNCIAS

ABE, J. M. *Fundamentos da lógica anotada.* Tese de Doutorado - Universidade de São Paulo, São Paulo: USP, 1992.

ABE, J. M.; PAPAVERO, N. *Teoria intuitiva de conjuntos.* São Paulo: Ed. Makron Books, 1992.

AITCHISON, J.; GILCHRIST, A. *Thesaurus Construction: a practical manual.* London: ASLIB, 1972.

BAEZA-YATES, Ricardo. *An extended model for full text databases.* Journal of the Brazilian Computer Society, vol.2, n°.3, April 1996.

BLUMER, A.; BLUMER J.; HAUSSLER D.; et al. *The smallest automaton recognizing the sub words of a text. Theoretical Computer Science,* n°. 40, p. 31-35, 1995.

BOOKSTEIN, A.; SWANSON, D. R. *Probabilistic models for automatic indexing.* J. American Society for Information Science, 25(5), pag. 312-18, 1974.

CALKINS, Mary L. *Free text or controlled vocabulary? Online*, v. 3, n. 2, p. 53-65, June, 1980.

CARROW, D.; NUGENT, J. *Comparison of free-text and index search abilities in an operating information system.* In: Information Management in the 1980's. New York. Proceedings... New York, v. 14, p. 232-238, 1977.

CARVALHO, André; BRAGA Antônio de P.; LUDERMIR, Teresa B. *Fundamentos de Redes Artificiais.* Rio de Janeiro: DCC/IM, COPPE / Sistemas, NCE/UFRJ, 1998. (*11° Escola de Computação*).

CHOUEKA, Y. *Looking for needles in a haystack OR locating interesting collocational expressions in large textual database*. Conference on User-Oriented Content-Based Text an Image Handling, MIT, Cambridge, Mass, pag 609-23, 1988.

CLEVERDON, C. W.; MILLS, J.; KEEN, M. *Factors determining the performance of indexing systems,* Vol. 1, Design, Vol. II, *Test Results,* ASLIB Cranfield Project, Cranfield, 1966.

COOL, C.; PARK, S.; BELKIN, N.J; et al. *Information seeking behavior in new searching environment.* CoLIS 2. Copenhagen. p. 403-416, 1996.

COOPER, W. S. *A definition of relevance for information retrieval. Inf. Storage Retrieval 7,* p. 19–37, 1971.

CRESTANI Fábio; LALMAS Mounia; RIJSBERGEN J. Van; et al. *"Is This Document Relevant?...Probably": A Survey of Probabilistic Models in Information Retrieval.* In ACM Computing Surveys, Vol. 30, No. 4, December, 1998.

CRESTANI, Fabio; van RIJSBERGEN, Cornelis J. A model for adaptative information retrieval. Journal of Intelligent Information Systems, v.8, 1997. p.29-56

CROFT, W. Bruce; JING Yufeng. *An association thesaurus for information retrieval.* University of Massachusetts at Amherst, Department of Computer Science. Amherst, MA 1003.

DA COSTA, N. C. A.; ABE Jair M.; DA SILVA Filho, João I.; et. al. *Lógica Paraconsistente Aplicada.* São Paulo: Editora Atlas, 1999.

DA COSTA, N. C. A.; ABE, J. M.; SUBRAHMANIAN, V. S. *Remarks on annotated logic,* Zeitschr. *f.* Math. *Logik und Grundlagen d. Math.,* 37: 561-

570, 1991.

DA COSTA, N. C. A.; *Ensaios sobre os fundamentos da lógica.* São Paulo: Hucitec, 1996.

DA COSTA, N. C. A.; SUBRAHMANIAN, V. S.; VAGO, C. *The paraconsistent logics* P. Zeitschr. *f.* Math. *Logik und Grundlagen d. Math.,* 37: 139-148, 1991.

DA COSTA, N.C.A.; MARCONI, D. *An overview of paraconsistent logic in the 80's,* In: *Logica Nova*, Berlin, 1987.

DA COSTA, N.C.A.; SUBRAHMANIAN, V.S.; HENSCHEN, L.J.; et al. *Automatic theorem proving in paraconsistent logic: Theory and implementation.* Estudos Avançados – nº. 03, 18p. Coleção de Documentos – USP – São Paulo, 1990.

DA COSTA, N.C.A.; SUBRAHMANIAN. V.S. *Paraconsistent logics as a formalism for reasoning about inconsistent knowledge bases.* Estudos Avançados – nº. 02, Coleção de Documentos, Série Lógica e Teoria da Ciência – USP – São Paulo, 1989.

DA SILVA Filho João I.; ABE Jair M.; MÁRIO, Maurício C.; et. al. *Conheça o Robô Emmy – o primeiro robô que funciona com lógica paraconsistente.* Revista Saber Eletrônica n.º 322, 1999.

DA SILVA Filho João I.; ABE Jair M. *Conheça a Lógica Paraconsistente.* Revista Saber Eletrônica nº. 317, 1999.

FORSYTH, R.; RADA R. *Machine learning – applications ins expert systems and information retrieval.* West Sussex, England: Ellis Horwood Series in Artificial Intelligence, 1986.

FOX, C. *A stop list for general text.* SIGIR Forum, 21(1-2), 19-35, 1990.

FOX, E. A.; NUTTER, J. T.; EVENS M.; et al. *Building a large thesaurus for information retrieval.* Paper present at the Second Conference on Applied Natural Language Processing. Association for Computational Linguistics, 101-08, 1988.

FRAKES, William B.; BAEZA-YATES, R. *Information Retrieval - Data Structures & Algorithms.* Prentice Hall PTR, 1992.

FURNAS W. George; DEERWESTER S.; DUMAIS T. S.; et al. *Information retrieval using a singular value decomposition model of latent semantic structure.* In Proc. of the 11th *Anual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pag. 465-480, 1988.

GLENN, B. T.; CHIGNELL, M. H. *Hypermedia: design for browsing.* In *Hartson, H.R, Hix, D. Advances in Human-Computer Interaction. Nerwood*, New Jersey, Ablex Publishing Corporation, v.3, 1992.

GUTHRIE, Louise et al. *The role of lexicons in natural language processing.* Communications of the ACM, v.39, n.1, Janeiro de 1996.

HARMAN, D. *Relevance feedback revisited.* In. *Proceedings of the 15th ACM SIGIR Conference,* Denmark, June 1992.

HARTER, S. P. *A probabilistic approach to automatic keyword indexing.* Parts I e II. J. American Society for Information Science, 26, pag. 197-206 and 280-89, 1975.

HENZLER, R. G. *Free or controlled vocabularies.* International Classification, v. 5, n. 1, p. 21-26, 32, Mar. 1978.

http://trec.nist.gov. Acessado em 01/02/2001.

http://www.w3.org. Acessado em 01/05/2001.

KLIR George J.; YUAN Bo. *Fuzzy Sets And Fuzzy Logic: Theory and applications.* New Jersey: Prentice Hall, 1995.

KORFHAGE, Robert. *Information Storage and Retrieval.* John Wiley & Sons, Inc., 1997.

LANCASTER, F. W. *Indexação e resumos: teoria e prática.* Brasília: Briquet de Lemos, 1993.

LANCASTER, F. W. *Information retrieval systems: characteristics, testing and evaluation.* 2nd, Ed. New York : Wiley, 1979.

LANCASTER, F. W. *Vocabulary control for information retrieval.* 2nd, Ed. Arlington: IRP, 1986.

LANCASTER, F. W.; FAYEN, E. G. *Information retrieval on-line.* Los Angeles: Melville, 1973.

LAPTEC'2000. *Anais I Congresso de Lógica Aplicada à Tecnologia, LAPTEC'2000,* 11 a 15 de setembro de 2000 / editor Jair M. Abe. São Paulo: Faculdade SENAC de Ciências Exatas e Tecnologia, 2000.

LIMA, Luciano R. S. de; LAENDER, Alberto H.F.; RIBEIRO-NETO, Berthier A. *Codificação Automática de Documentos em Bases de Dados Médicas: Um Estudo Comparativo. in XIII SBBD Simpósio Brasileiro de Banco de Dados.* Maringá: Ideal, Outubro de 1998.

MARON, M. E.; KUHNS, J. L. *On relevance, probabilistic indexing and retrieval.* J. ACM 7, 216-244, USA, 1960.

McGILL, M. et al. *An evaluation of factors affecting document ranking by information retrieval systems.* Project report. Syracuse, New York: Syracuse

University School of Information Studies, 1979.

MENDONÇA, Gustavo C. G.; SILVA Ilmério R. da ; RIBEIRO-NETO, Berthier A. *Uma Técnica de Filtragem para Recuperação de Informação. in XIII SBBD Simpósio Brasileiro de Banco de Dados.* Maringá: Ideal, Outubro de 1998.

MILLER, G. A. , BECKWITH R., FELLBAUM C., et al. *Introduction to Wordnet: Na On-line Lexical Database*. Technical report, August, 1993. Disponível em http://www.cogsci.princeton.edu/~wn1993.

MINKER, JACK; WILSON, GERALD A.; ZIMMERMAN BARBARA H. *An evaluation of query expansion by the addition of clustered terms for a document retrieval system*, IP&M, Vol. 8, 329-348, 1972.

MIZZARO, S. *Relevance: The whole (hi)story*. Tech. Rep. UDMI/12/96/RR (Dec.), Diparti-mento di Matematica e Informatica, Universita' di Udine, Italy, 1996.

OGAWA, Y; MORITA, T; KOBAYASHI, K. *A fuzzy document retrieval system using the keyword connections matrix and a learning method.* Fuzzy Sets and Systems, 39: 163-179, 1991.

PEARL, Judea. *Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference.* revised 2nd printing Addison-Wesley, 1991.

QIU, YONGGANG; FREI, H. P. *Concept based query expansion.* SIGIR'93, 160-169, 1993.

RAITT, D. J. *Recall and precision devices in interactive bibliographic search and retrieval systems.* Aslib Proceedings, v. 32, n. 7/8, p. 281-301, July/Aug. 1980.

RIBEIRO-NETO, Berthier A.; YATES-BAEZA, Ricardo. *Modern Information Retrieval.* New York: ACM Press Book, 1999.

RIBEIRO-NETO, Berthier A; MUNTZ Richard. *A belief network model for IR.* In Proc. of the 19th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval. pages 253-260. Zurich. Switzerland. 1996.

RIJSBERGEN C. J. van. *Information Retrieval.* London, second edition, 1979.

RIJSBERGEN C. J. van; HARPER, D. J.; PORTER M. F. *The selection of good search terms.* IP&M, Vol. 17, 77-91, 1981.

ROCCHIO, J. J. *Relevance feedback in information retrieval.* In: The SMART retrieval system: experiments in automatic document processing. G. Salton, ed. pp 313-323 Prentice-Hall.

ROGET, Peter; *Roget's II The New Thesaurus.* Houghton Mifflin Company, Boston, USA, 1998.

SALTON Gerard; FOX Edward A.; WU Harry. *Extended Boolean information retrieval.* Communications of the ACM, 26(11):1022-1036, November 1983.

SALTON Gerard; McGILL J. M. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Co., New York, 1983.

SALTON, G. *Automatic information organization and retrieval.* McGraw-Hill Book Company, 1968.

SALTON, G.; BUCKLEY, C.; YU, C. T. *An evaluation of term dependence models in information retrieval.* LNCS 146, 151-173, 1983.

SALTON, G.; YANG, C. S. *On the specification of term values in automatic indexing.* Journal of Documentations, 29(4), pag. 351-72, 1973.

SALTON, Gerard; BUCKLEY, Christopher. *Improving retrieval performance by relevance feedback.* In: SPARCK-JONES, e WILLET, 1997.

SENKO, M.E. *Information storage and retrieval systems.* In *Advances in Information Systems Science,* (Edited by J. Tou) Plenum Press, New York, 1969.

SERACEVIC, T. *The concept of "relevance" in information science: A historical review. In Introduction to Information Science,* T. Seracevic. Ed., R. Bower, New York, Chapter 14, 1970.

SILBERSCHATZ Abraham; KORTH Henry F.; SUDARSHAN S. *Sistemas de Banco de Dados.* 3. ed. São Paulo: Makron Books, 1999.

SILVA, EDNA L. da; MENEZES ESTERA M.; *Metodologia da pesquisa e elaboração de dissertação.* Laboratório de Ensino a Distância da UFSC. Florianópolis, 2000.

SILVA, Ilmério R. da; RIBEIRO-NETO, Berthier A. *Avaliação de Desempenho de um Modelo Bayesiano para Recuperação de Informações em Bibliotecas Digitais. in XIII SBBD Simpósio Brasileiro de Banco de Dados.* Maringá: Ideal, Outubro de 1998.

SPARCK-JONES, Karen; WILLET, Peter (eds). *Readings in Information Retrieval.* San Francisco: Morgan Kaufmann, 1997.

SPARK JONES, K. *Automatic keyword classification for information retrieval,* Butterworth, 1971.

SPARK JONES, K.; JACKSON, D. M. *The use of automatically-obtained keyword classifications for information retrieval,* IP&M, Vol. 5, 175-201, 1970.

SPARK JONES, K; NEEDHAM, R. M. *Automatic term classification and retrieval,* IP&M, Vol. 4, 91-100, 1968.

SPINK A.; WILSON T. D. *Toward a theoretical framework for information*

retrieval (IR) evaluation in an information seeking context. In: Mira '99. eletronic Workshops in Computing. S. Draper, M. Dunlop, I. Ruthven and C. J. van Rijsbergen (ed).

SRINIVASAN, P. *A comparison of two-poisson, inverse document frequency and discrimination value models of document representation.* Information Processing and Management, 26(2) pag. 269-78, 1990.

SUBRAHMANIAN, V. S. *On the semantics of quantitative logic programs. Proc. 4th IEEE Symposium on Logic Programming, Computer Society Press,* Washington D. C., p. 173-182, 1987.

SVENONIOUS, E. *Design of controlled vocabularies.* In: Encyclopedia of Library and Information Science. New York : Marcel Dekker, v. 45, p. .82-108, 1988.

SVENONIOUS, E. *Natural language vs controlled vocabulary.* In: Canadian Conference on Information Science, Ontario. Proceedings... [S. l.: s. n.] p. 141-150, 1976.

SVENONIOUS, E. *The intellectual foundation of information organization.* Cambridge : MIT, 2000.

TURTLE, Howard R. *Inference networks for document retrieval.* Tech. rep., University of Massachusetts Ph.D. dissertation, 1991.

TURTLE, Howard; CROFT, W. Bruce. *Efficient probabilistic inference for text retrieval.* In *Proceedings of RIAO 3*, 1991.

TURTLE, Howard; CROFT, W. Bruce. *Evaluation of an inference network-based retrieval model.* ACM Transactions on Information Systems, 9(3):187-222, July 1991.

TURTLE, Howard; CROFT, W. Bruce. *Inference networks for document retrieval.*

In Proc. of the 13th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval , pages 1-24, Brussels, Belgium, 1990.

WANG, Y-C; VANDENDORPE, J.; and EVENS M. *Relationship thesauri in information retrieval.* J. American Society of Information Science, p. 15-27, 1985.

WATERWORTH, J. A.; CHIGNELL, M. H. *A manifesto for hypermedia usability research.* Hypermedia 1, p. 205-234, 1989.

WILKINSON, R.; HINGSTON, P. *Using the cosine measure in a neural network for document retrieval.* In Proc. *of the ACM SIGIR Conference on Research and Development in Information Retrieval,* pag. 202-210, Chicago, USA, Oct 1991.

PEREZ, Ernest. *Text enhancement: controlled vocabularies vs free text.* Special Libraries, v. 72, n. 3/4, p. 183-192, 1982.