RECUPERAÇÃO DE INFORMAÇÃO USANDO COMPUTAÇÃO NEBULOSA A PARTIR DE DOCUMENTOS COM ESTRUTURAS HETEROGÊNEAS

Gustavo Alberto Giménez Lugo [⊗]
Escola Politécnica
Universidade de São Paulo
São Paulo - SP
gustavo@pcs.usp.br

Marco Túlio Carvalho de Andrade
Escola Politécnica
Universidade de São Paulo
São Paulo - SP
mtulio@pcs.usp.br

Jaime Simão Sichman
Escola Politécnica
Universidade de São Paulo
São Paulo - SP
jaime@pcs.usp.br

RESUMO

Este trabalho apresenta um experimento que usa técnicas nebulosas para implementar um sistema protótipo para recuperação de informação, reduzindo de n para 3 o número de seções consideradas em um documento. O objetivo foi verificar o comportamento e o desempenho de tal enfoque quando aplicado a uma coleção de documentos com estruturas heterogêneas, a priori desconhecidas, articulando o texto. Os resultados foram promissores apesar da simplicidade do enfoque, criando um grande interesse na exploração de modelos mais sofisticados de recuperação de informação usando a teoria nebulosa para a construção de ontologias particulares.

ABSTRACT

This work presents an experiment using fuzzy techniques to implement a prototype system for information retrieval, reducing from n to 3 the number of sections considered in a document corpora. The goal was to verify the behavior and performance of such an approach when applied to a collection of documents with heterogenous, a priori unknown, structures articulating the text. The results were promising, in spite of the rather primitive approach, instillating a deep interest in exploring a more sophisticated information retrieval model, better exploiting fuzzy theory as a way to build particular ontologies.

1 INTRODUÇÃO

A problemática de explosão da informação, hoje tão visada pela pesquisa, já era apontada por Vannevar Bush na década de 40 (Bush, 1945 apud Saracevic, 1995) e é em volta deste ponto que gira o próprio núcleo da recuperação de informação.

A conceituação de Recuperação de Informação, definida inicialmente por Calvin Mooers (Mooers, 1951 apud Saracevic, 1995), vem dada, em geral (Frakes, 1992; Russell e Norvig, 1995; Turtle e Croft, 1997), de uma forma eminentemente funcional e não descritiva:

> *Recuperação de informação é o nome do processo ou método onde um posssível usuário de informação pode converter a sua necessidade de informação numa lista real de citações de documentos armazenados que contenham informações úteis a ele . . . recuperação de informação abarca os aspectos intelectuais da descrição da informação e a sua especificação para busca, assim como também quaisquer sistemas, técnicas ou máquinas que sejam empregadas para efetuar a operação (Mooers, 1951 apud Saracevic, 1995).*

De qualquer modo, uma vez de posse dos dados da consulta são selecionados os documentos que se apresentam como mais relevantes comparando a consulta com representações dos documentos previamente armazenadas, opcionalmente revisando a representação da consulta para tratar consultas posteriores. O processo de recuperação envolve assim a aquisição da consulta, seleção dos documentos e revisão da consulta. Uma coleção de documentos pode incluir documentos textuais e não textuais, tais como fotografias, som, vídeo, módulos de software, etc.

A explosão de informação é um complexo problema social, cognitivo, cultural e de comunicação e não simplesmente um problema técnico. Isto traz a tona a visão de Zadeh (Zadeh, 1965; Zadeh,1973) quanto às vantagens de usar um tratamento nebuloso para lidar com sistemas que interagem com seres humanos. Muito mais ainda quando a própria essência da computação nebulosa está centrada no universo do discurso lingüístico humano, enquanto aquisição e representação do conhecimento. Autores como (Saracevic, 1995) chegam a expressar o sucesso dos sistemas de recuperação de informação em função do problema:

> *O quão satisfatório foi e é a Recuperação de Informação pode ser medido pela aplicação na resolução do problema de explosão da informação nas áreas em que foi aplicado?*

Isto leva a outras questões correlatas, como:

> *Quão satisfatoriamente a Recuperação de Informação fornece suporte às pessoas nas situações em que elas se defrontam com problemas de pesquisa, achado, uso e interação com informação advinda da massa de informações existentes e a miríade de escolhas possíveis?*

Recuperação de informação é aplicada, dentre outras formas, na busca e navegação através da Internet e constitui o coração da pesquisa e o desenvolvimento das bibliotecas digitais.

Qualquer sistema de recuperação de informação inclui três componentes básicos: identificação e representação do conteúdo do documento, aquisição e representação da necessidade de informação e a especificação da função de comparação que seleciona os documentos relevantes baseado nas representações. A incerteza aparece em cada um dos componentes citados (Croft e Turtle, 1992; Turtle e Croft, 1997).

[⊗] Bolsista CNPq

Um aparente paradoxo é subjacente ao domínio dos sistemas de informação em geral, desde que tais sistemas devem manipular informações permeadas de incertezas (ou seja, informações sobre as quais não se possui conhecimento em graus determinados ou não) sendo os sistemas, eles próprios, recheados de possíveis fontes de incerteza (Mamdani, 1997). Deve ser notado que o termo pode ser usado tanto num contexto mais amplo definindo uma "informação imperfeita", como em outro mais específico para descrever uma forma particular de imperfeição. No primeiro caso o termo refere-se a situações onde a informação disponível "carece de perfeição" (Motro, 1997), no segundo descreve casos onde a correção da informação disponível está em dúvida.

Considerando um sistema de informação *D* que tenta representar uma fração do mundo real *W*, *D* é *bem fundado* (*sound*) se *D está contido* em *W*; ou seja, a informação armazenada *contem* somente informação *verdadeira*. *D* é *completo* se *contém W*; ou seja, a informação armazenada inclui *toda* a informação que o sistema presume modelar. Em outras palavras um sistema de informação é perfeito se inclui *toda* a verdade e *nada além* da verdade.

No caso de um sistema de informação não ser *bem fundado* e *completo*, uma alternativa consiste em melhorar as técnicas de modelagem usadas na *aproximação* de *W*. Seja para poder representar informação *imprecisa* ou tratar *incerteza*.

Considerando somente corpos textuais, convém lembrar que a mensagem que eles carregam permeia as funções da linguagem (Chalhub, 1999) e está permeada delas. Tais funções indicam:

o quê - função referencial;
quem - função emocional;
para quem -função conativa;
onde - função fática: que indica o canal;
como - função poética;
o código usado e reproduzido- função metalinguística.

Estes fatores articulam-se no texto (Guimarães, 1999) segundo convenções particulares das quais se pode tirar proveito na determinação do que vem a ser um texto relevante para o usuário.

As formas de estruturação do texto podem ser descritivas, narrativas, dissertativas ou combinações delas. Estas formas manifestam-se no tipo de integração das partes temáticas ou estruturais do texto, sempre levando em conta início e fim.

É fácil perceber então que a informação não está uniformemente distribuída num texto, uma premissa de muitos sistemas de recuperação de informação. (Bordogna e Pasi, 1997) mostraram de forma convincente o quanto se pode ganhar ainda aplicando computação nebulosa na representação de documentos uniformemente estruturados atrelados a formas diferenciadas de pesar a relevância.

Como contraposição às condições trabalhadas pelas pesquisadoras, porém, aparecem com enorme frequência situações onde se deve discernir sobre documentos de estruturas as mais variadas. É este o leitmotiv do presente trabalho, que não se pretende inovador como (Bordogna e Pasi, 1997) mas sim exploratório da computação nebulosa como meio de representação de um certo conhecimento. Neste caso, os documentos tratados, mesmo possuindo estruturas diferentes, partilham de um modo de construção de texto incorporado hoje à escrita até coloquial na nossa cultura - possuir um início um meio e um fim. Partes estas de tamanhos e relacionamentos os mais diversos, mas que podem ser usadas futuramente como âncoras de um tratamento lingüístico mais refinado que reflita e se apoie na incerteza do que é pedido e do representado, não apenas usando uma linha de análise cegamente precisa.

A sequência do texto discorre sobre a modelagem do enfoque utilizado, detalhando a representação nebulosa dos documentos e a estrutura computacional que permite a sua manipulação, a representação das necessidades de informação do usuário e a escolha da forma de quantização dos escores dos documentos que permitem julgar a sua relevância para o usuário. Finalmente são discutidos detalhes sobre a implementação do protótipo e finalmente os resultados atingidos através dele.

2 UM ENFOQUE DE RECUPERAÇÃO DE INFORMAÇÃO BASEADO EM COMPUTAÇÃO NEBULOSA

Baseado no exposto na seção anterior partiu-se para a construção de um sistema protótipo de forma a explorar o comportamento de recuperação usando uma representação nebulosa dos documentos a serem indexados. O sistema pode ser esquematicamente representado como no modelo da figura 1.

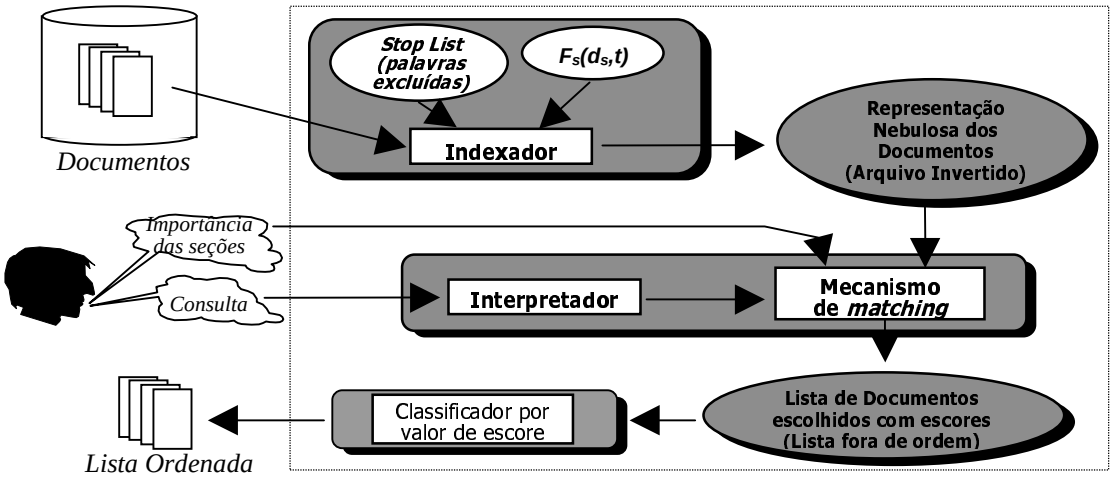

Fig. 1 - Representação esquemática do protótipo de sistema de recuperação de informação baseado em representação nebulosa de documentos.

A seguir serão descritos tanto a representação nebulosa usada quanto detalhes de implementação. Estes detalhes dizem respeito ao tipo de estrutura de dados empregado na representação dos corpos textuais e ao interpretador simples usado para tratar as consultas do usuário.

2.1 Representação Fuzzy de Documentos Parcialmente Estruturados.

Uma premissa do experimento é que os arquivos tratados não partilham da mesma estrutura de articulação do texto. Cada par termo/documento *(t, d)* não está associado simplesmente a um peso *F(d,t)*, mas a um conjunto de valores $F_1(d,t)$, … $F_i(d,t)$, …, $F_n(d,t)$, denotando o grau de significação do termo *t* na seção *i* (neste caso *i* foi escolhido como tendo três valores possíveis, correspondentes a: Início, Meio e Fim).

Quanto ao critério de relevância para o usuário, ele baseia-se na linha iniciada por Kent et al. (Kent et alii, 1955 apud Saracevic, 1995) que foram os primeiros a propor o critério de relevância e as medidas de precisão e relevância (mais tarde renomeado *recall* - recuperação) como métricas plausíveis do processo de recuperação. O valor de recall indica quantos documentos foram recuperados dentre o total considerado relevante na coleção pesquisada. A precisão ilustra qual a proporção entre os documentos considerados relevantes e aqueles efetivamente recuperados.

O grau de significação final resultante *F(d,t)* é computado combinando os graus de significação parciais $F_i(d,t)$ através de uma função especificada levando em conta a escolha pelo usuário da seção mais relevante. Esta função é identificada por uma variável lingüística representando a *posição* no texto a ser privilegiada, com os seguintes valores: início, meio e fim. Como caso particular é permitido ao usuário levar em conta todas as seções de forma conjunta.

A função $F_{i=1,2,3}$*: D x T* é definida para cada uma das seções artificialmente estabelecidas segundo a fórmula dada pela equação (1):

documentos indexados pelo termo procurado traduz a intenção de realçar o valor de termos que aparecem mais raramente nos documentos, ou seja, aqueles que ajudam a discriminar mais fortemente um documento que o contem.

Pode-se observar que a equação (1) pode não produzir somente resultados no intervalo nebuloso clássico [0,1]. Isto, porém, não afeta de maneira decisiva o resultado pois o objetivo central é obter uma forma de pontuação que permita ordenar os documentos por grau de relevância.

As funções de pertinência simples $F_i(d_i, t)$ são mostradas na figura 2 e são fartamente usadas em aplicações de computação nebulosa (Von Altrock, 1995), dispensando maiores comentários.

Diferentemente do trabalho de Bordogna e Pasi (Bordogna e Pasi, 1997) que diferencia seções estruturadas e não estruturadas, as seções são tratadas apenas levando em conta as funções de pertinência simples representadas na figura 2 como visto a seguir:

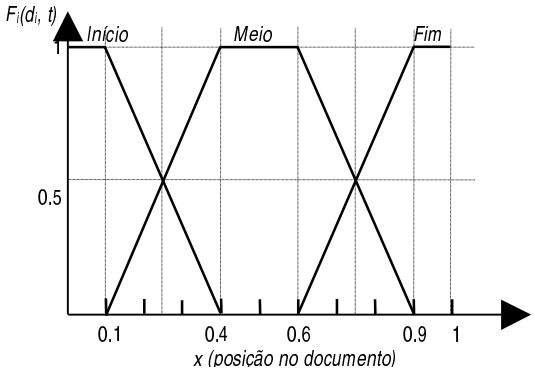

Fig. 2 - Funções de pertinência para etiquetas Início Meio e Fim da variável posição.

A posição de uma instância de um termo no texto é determinada como sendo a razão simples entre o número que indica a posição da instância do termo no documento e o número total de palavras existentes no documento.

$$F_i(d, t) = \frac{\sum_{k=1}^{N^o \text{ de ocorrências de t}} F_i(d_i, t_k)}{\sum_{j=1}^{N^o \text{ de ocorrências de tt}} F_i(d_i, tt_j)} \bullet \log\left(\frac{N_{total\ documentos}}{NDT_{total\ docs\ indexados\ por\ t}}\right) \qquad \text{Eq. (1)}$$

onde:

$F_i(d, t)$ = valor da função de pertinência resultante para o termo t na seção i do documento d.
d = documento a ser indexado
d_i = seção i do documento d
t = termo a ser qualificado
tt = termo que aparece com maior freqüência na seção do documento, usado como referência.
$F_i(d_i, t_k)$ = função de pertinência para uma instância k de um termo t na seção i do documento

A primeira parte da equação (1), dada pela razão dos somatórios, pretende medir a intensidade com a qual aparece a representação do conceito sumarizado num termo qualquer.

Parece apropriado, portanto, medir a quantidade de informação relativa do termo procurado comparado ao termo que aparece com maior frequência na porção de texto considerada. Este tipo de medida é consagrado em computação nebulosa enquanto ligado ao conceito de entropia (De Luca e Termini, 1972; Trillas e Riera, 1978).

Na segunda parte da equação (1), o logaritmo da razão entre o número total de documentos e o número de

2.2. Estrutura de Dados Usada na Representação

O sistema implementado para efetivar o experimento concretiza a representação nebulosa dos documentos através dos dados armazenados numa estrutura tipicamente usada em recuperação de informação, denominada arquivo invertido.

O arquivo invertido guarda as palavras que foram indexadas. Junto com cada palavra existe uma lista de documentos nos quais o termo aparece. Associados a cada documento são armazenados dois vetores, um correspondente aos valores computados para o termo em

cada seção do documento correspondente e o outro aos valores usados para normalização, usando os valores computados para os termos que aparecem com mais freqüência em cada seção. O número de documentos indexados por cada palavra é um dado obtido de forma direta desta lista.

Este arquivo invertido foi implementado usando um *trie* (*tree retrieval*) (Ammeraal, 1996) cuja representação pode ser vista na figura 3. Nas folhas da árvore construída são guardados o termo referenciado, algumas variáveis auxiliares para o pré-processamento e um ponteiro para a lista de documentos que contém o termo.

Em cada nó da lista, além do nome do documento, são guardados os valores dos somatórios de $F_i(d_i, t)$ e de $F_i(d_i, tt)$ para as seções correspondentes, calculados pelo indexador durante o pré-processamento. A escolha do trie foi arbitrária, podendo muito bem a estrutura estar baseada numa árvore-B.

A construção de dita estrutura permite um acesso rápido às listas de documentos relevantes, tendo como custo o tempo de pré-processamento.

Para a indexação, todas as letras do alfabeto português foram incluídas, além dos acentos correspondentes e da aceitação do traço (-) como elemento de composição de termos.

Na versão atual o protótipo não permite a busca de frases tratadas como termos compostos, sendo, porém, esta alternativa de implementação relativamente imediata.

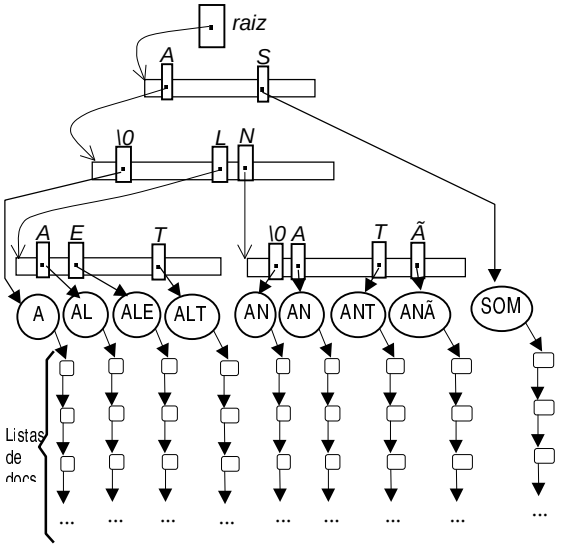

Fig. 3 - Arquivo invertido implementado usando listas de documentos dependuradas das folhas de um trie.

Foi elaborada uma lista de palavras não indexáveis (também chamada de *stop-list*) de forma tal que somente palavras com mais de uma letra e não pertencentes ao conjunto *S* são consideradas:

S = { AO, AS, DA, DE, DO, DOS, DAS, EM, COMO, QUE, NA, NO, OS}

A *stop-list* aparece na tentativa de evitar a indexação de termos com baixo poder de discriminação. Esta lista pode ser facilmente alterada, de acordo ao universo lingüístico do domínio focalizado. No entanto, as palavras que compõem a coleção foram escolhidas sem levar em conta qualquer domínio específico do conhecimento.

2.3. Representação da Consulta do Usuário

Das diversas formas possíveis de representação de consultas do usuário foi escolhida a consulta Booleana, já tratada de forma específica na literatura (Kacprzyc e Zadrozny, 1997; Kostelansky, 1998). Muito embora ciente das suas limitações, considerou-se suficiente no contexto desta experiência.

A necessidade de informação do usuário é então expressada por ele na forma de uma consulta que é interpretada no protótipo, sendo o resultado levado em conta (junto com a escolha da seção feita pelo usuário) no mecanismo de matching.

Foi adicionada uma alternativa (implementada por <valorImportância>) de forma a permitir ao usuário restringir a resposta somente aos documentos mais relevantes, ou seja, com pontuação acima de patamares específicos. As formas aceitas podem ser ilustradas usando uma gramática análoga à BNF (*Backus-Naur Form* - Forma de Backus-Naur) para representação das regras de produção que descrevem a sintaxe:

<consulta>:= <abreChave> <seqüência_de_condições> <fechaChave>
<seqüência_de_condições>:= <seqüência_subCondições>
| <seqüência_subCondições> OU <seqüência_de_condições>
<seqüência_subCondições>:= <subCondição> | <subCondição>E <seq_subCondições>
<subcondição>:= <abreParentesis> <termo> <vírgula> <valorImportância><fechaParentesis>
<valorImportância>:= IMPORTANTE | MAIS_OU_MENOS | POUCO_IMPORTANTE

onde:

<termo> é uma palavra procurada

<valorImportância> é um parâmetro opcional que indica o grau de importância mínima que deve ter um documento para ser considerado, sendo a semântica adotada a que segue: quanto maior for o valor de importância que acompanha o termo, somente os documentos cujos escores estiverem acima de determinados valores de corte serão considerados. Ou seja, se o usuário escolher POUCO_IMPORTANTE, equivale a pedir "mostre *todos* os documentos achados *inclusive* aqueles com escores mais baixos". O valor por defeito (*default*) é POUCO_IMPORTANTE.

Um exemplo de consulta bem formada é a seguinte:

{ (ALTO , IMPORTANTE) E (ANÃO , IMPORTANTE) OU (SOMA) }

onde as palavras a serem procuradas são ALTO, ANÃO e SOMA e o grau de importância mínimo estabelecido para a palavra SOMA é aquele definido por *default*, i.e., POUCO_IMPORTANTE.

2.4. Assignação de Escore a um Documento

Como saída, um sistema de recuperação de informação deve fornecer uma lista ordenada de documentos relevantes. Isto é devido ao fato de que a ordem em que os documentos são apresentados ao usuário é notadamente relevante no juízo e no uso que o usuário fará da lista obtida.

No modelo adotado, a contrução de um valor final único de escore F(d, t) para um documento *d* em relação a um termo *t* leva em consideração os seguintes fatores:

I. O valor da função $F_i(d, t)$, ou seja o valor da função de pertinência resultante para o termo *t* na seção *i* do documento *d*, tal como definido na equação (1).
II. O valor escolhido pelo usuário para a variável lingüística *posição*: início, meio ou fim.

Este conjunto de parâmetros é integrado segundo a fórmula:

$$F(d, t) = \sum_{i=1,\ldots,3} (a_i \bullet F_i(d, t)) \qquad \text{Eq. (2)}$$

onde:

a_i *: depende do valor escolhido para a variável lingüística posição, e.g., se o valor fornecido pelo usuário for INICIO, então* $a_1 = 1$ *e* $a_i = 0$ *para qualquer outro valor de i. Se o valor não for fornecido, assume-se* $a_i=1 \ \forall i$*. Este valor é solicitado para cada consulta.*

$F_i(d, t)$ *: é o valor da função de pertinência resultante para a seção i.*

Outras formas de agregação são possíveis e evidentemente de desejável tratamento. A apresentada na equação (2) foi escolhida pela sua simplicidade e pelo fato de poder ser usada como referencial em relação a outras mais sofisticadas.

3 RESULTADOS

Como primeiro passo foi escolhido um conjunto de trinta documentos escolhidos por um especialista, dentre os textos produzidos por ele mesmo, através de vários anos de trabalho.

A partir daí foram executados os seguintes passos:

- o próprio especialista classificou os documentos escolhidos em cinco grandes classes de tópicos de interesse;
- para cada classe estabelecida era necessário um julgamento de relevância mais específico, portanto os documentos foram novamente avaliados pelo especialista em graus de relevância em relação ao tópico;
- uma família de palavras foi escolhida pelo especialista como sendo representativa de cada classe de conceito envolvido;
- o especialista elaborou um conjunto de consultas dividido por classes que deveriam retornar os documentos relevantes para cada tópico sem especializar a consulta por seções nos documentos, apenas levando em conta a definição de consulta bem formada aceita pelo interpretador. Estas consultas deveriam servir de referência para comparação com os resultados atingidos usando a especialização por seções nos documentos;
- uma outra coleção de consultas foi elaborada pelo especialista, desta vez dirigida às seções dos documentos.

A coleção de documentos ocupou um espaço de aproximadamente 1,1 Mb, na forma de arquivos Word e não puramente textuais, cada texto contendo em média aproximadamente 3.000 palavras. Isto evidentemente tende a degradar os resultados obtidos com o protótipo, mas a escolha foi feita para aproximar o contexto das condições de uso do especialista.

Quanto ao protótipo em si, implementado segundo os detalhes mostrados na seção anterior, foi desenvolvido usando o compilador Microsoft Visual C++ 2.0, sendo estruturado em várias classes hierarquicamente integradas, executando no sistema operacional Windows 95, inicialmente em um computador 486 DX4-100.

O espaço total ocupado pelas representações em tempo de execução foi de aproximadamente 20 Mb. Muito embora o tempo de indexação total da coleção fosse em torno de 30 minutos, o tempo de resposta médio para as consultas foi em torno de 1 segundo, isto no 486. Com posterioridade, o mesmo teste foi repetido usando um Pentium III, 550Mhz, tendo levado apenas 40 segundos para indexar o mesmo corpo de documentos.

Deve-se chamar a atenção sobre o fato de não ter sido executada nenhuma tarefa de refinamento do código ou dos algoritmos, portanto estes tempos e espaços envolvidos possuem larga faixa para serem minimizados.

Os resultados das consultas pré-elaboradas, considerando os juízos de relevância do especialista, são comparados nas tabelas 1 e 2, usando as métricas clássicas de precisão e recall, definidas na seção 2.1.

Tabela 1. Resultados medidos das consultas considerando todo o texto. A cada classe corresponde uma consulta, cinco ao todo.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Precisão	60 %	50 % (*)	100%	75%	43% (♣)
Recall	60%	17% (*)	25%	17%	75% (♣)

Tabela 2. Resultados medidos das consultas focalizando seções do texto. A cada classe corresponde uma consulta, exceto na classe 2 - seis ao todo.

	Classe 1	Classe 2		Classe 3	Classe 4	Classe 5
Precisão	60%	100% (*)	33% (**)	33%	0% (♣♣)	38%
Recall	60%	17% (*)	33% (**)	8%	0% (♣♣)	75%

Observações:

(*) *Um documento era parcialmente relevante, porém, a relação foi descoberta somente observando o resultado da consulta*

(**) *A relevância dos documentos foi ressaltada com a busca por seções*

(♣) *Um documento "não relevante" foi "descoberto" como sendo relevante após observar o resultado da consulta.*

(♣♣) *A consulta foi julgada inadequada pelo especialista após observar o resultado da consulta*

4 CONCLUSÃO

Parece natural, à vista dos resultados obtidos, que as diversas formas de tratamento da entropia em computação nebulosa (De Luca e Termini, 1972; Trillas e Riera, 1978) podem ser melhor aproveitadas na representação e recuperação de informação, da mesma forma, através do refinamento da representação pode-se realçar os melhores resultados usando critérios de difusão/intensificação (Friedhover e Simões, 1996), quando da obtenção de um resultado final único para cada documento.

Acredita-se, para além dos resultados obtidos com esta experiência, que a computação nebulosa tem muito mais a oferecer ao domínio da recuperação de informação do que simplesmente extender as fronteiras dos métodos já conhecidos. Isto deve ser reforçado pois nela baseiam-se críticas apontando as "limitações" da computação nebulosa aplicada à recuperação de informação, tais como (Turtle e Croft, 1997):

1. *os modelos suportam bem a ponderação de termos mas não modelam diretamente o uso de pesos para as consultas;*

2. *documentos são geralmente ordenados usando um número pequeno de termos; ainda, um documento contendo a ocorrência de um só termo de uma consulta com função de pertinência igual a 1 é classificado com a mesma importância de um outro que contem todos os termos da consulta com valor igual a 1;*

3. *resultados diferentes podem ser atingidos para consultas logicamente equivalentes.*

As citadas críticas provém de dois pesquisadores considerados como grandes vultos no âmbito da recuperação de informação. As tais "limitações" advém em parte do fato de ser usado o modelo de consulta Booleano. E quanto a elas propriamente, as que convém explorar são a segunda, que, como demonstra o próprio enfoque adotado neste trabalho, pode ser contornada de forma fácil usando medidas relacionadas à entropia; e a terceira, que não é um demérito mas sim uma virtude a ser explorada com ferramentas adequadas (não necessariamente a lógica tradicional) desde que o próprio ponto de vista do usuário quanto ao conteúdo de conhecimento sendo acessado não pode, e não deve ser tratado através de um filtro único, deve-se sim adequar o sistema como forma de mediação não limitada por simples preconceitos metodológicos ou tecnológicos.

No caso de sistemas baseados em texto, parece plausível a elaboração de ontologias nebulosas de forma a refletir com muito mais precisão as categorias do mundo real nos diferentes universos de discurso partilhados por comunidades de usuários. Agentes que referenciem estas ontologias aparecem como sendo muito mais eficientes do que aqueles cujo conhecimento linguístico é extendido apenas com meros dicionários de sinônimos e inclusive metadados. Como vantagem adicional, a computação nebulosa fornece os meios para estabelecer marcos semânticos baseados em expressões linguísticas muito mais intuitivas que construções meramente probabilísticas.

REFERÊNCIAS BIBLIOGRÁFICAS

AMMERAAL, L. *Algorithms and Data Structures in C++*. England: John Wiley&Sons, 1996.

BORDOGNA, G. and PASI, G. A Fuzzy Information Retrieval System Handling User's Preferences on Document Sections. In: FUZZY INFORMATION ENGINEERING - A GUIDED TOUR OF APPLICATIONS. USA: John Wiley & Sons, Inc.; Dubois, D., Prade, H. and Yager, R.R. (eds.);1997. P. 265-281.

BUSH, V. As we may think. *Atlantic Monthly*, USA, v.176, n.1, p. 101-108, 1945.

CHALHUB, S. *Funções da Linguagem*, 10ª edição. São Paulo: Editora Ática, 1999.

CROFT, W.B. and TURTLE, H.R. Text Retrieval and Inference. In: TEXT-BASED INTELLIGENT SYSTEMS: CURRENT RESEARCH AND PRACTICE IN INFORMATION EXTRACTION AND RETRIEVAL. USA: Lawrence Erlbaum Associates, Publishers, Jacobs, P.S. (ed.), 1992, p. 127-155.

DE LUCA, A.. and TERMINI, S. A definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory. *Information and Control*, v.20, p.301-312, 1972.

FRAKES, W.B. Introduction to Information Storage and Retrieval Systems. In: INFORMATION RETRIEVAL-DATA STRUCTURES & ALGORITHMS. USA: Prentice Hall; Frakes, W.B. and Baeza-Yates, R (eds.), 1992, p.1-12.

FRIEDHOVER, M., SIMÕES, M. G. Fuzzy decision making: grau de compensação em atribuição de importâncias e agregação de valores. In: SEMINÁRIO DE CONTROLE E MODELAMENTO INTELIGENTE, 1, 1996. São Paulo: Epmc-Las, 1996. P.10-4.

GUIMARÃES, E. *A Articulação do Texto*, 7ª ed. São Paulo: Editora Ática, 1999.

KACPRZYC, J. and ZADROZNY, S. Fuzzy queries in Microsoft Access V.2. In: FUZZY INFORMATION ENGINEERING - A GUIDED TOUR OF APPLICATIONS. USA: John Wiley & Sons, Inc.; Dubois, D., Prade, H. and Yager, R.R. (eds.); 1997. P. 223-232.

KENT, A., BERRY, M., LEUHRS, F.U., and PERRY, J.W. Machine literature learning VIII. Operational criteria for designing information retrieval systems. *American Documentation*, USA, v.6, n.2, p.93-101, 1955.

KOSTELANSKY, J. *Fuzzy Information Retrieval or a way to more Intelligent and Human Querying*. February, 1998. Material disponível na Internet. *http://www.kredit.sk/DB-fuzzy/fuzzy_info.htm.*

MAMDANI, E.H. On the classification of uncertainty techniques in relation to the application needs. In: UNCERTAINTY MANAGEMENT IN INFORMATION SYSTEMS - FROM NEEDS TO SOLUTIONS. USA: Kluwer Academic Publishers; Motro, A. and Smets P. (eds.); 1997. P.397-411.

MOOERS, C. N. Zatocoding applied to mechanical organization of knowledge. *American Documentation*, USA, v.2, p.20-32, 1951.

MOTRO, A. Sources of uncertainty, imprecision, and inconsistency in information systems. In: UNCERTAINTY MANAGEMENT IN INFORMATION SYSTEMS - FROM NEEDS TO SOLUTIONS. USA: Kluwer Academic Publishers; Motro, A. and Smets P. (eds.); 1997. P.9-34.

RUSSELL, S. J. and NORVIG, P. *Artificial Intelligence*. USA: Prentice Hall, 1995.

SARACEVIC, T. Evaluation of Evaluation in Information Retrieval. In: CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. 18TH ANNUAL INTERNATIONAL SIGIR. Seattle, USA. Proceedings... USA: ACM Press, 1995. P. 137-146.

TRILLAS, E. and RIERA, T. Entropies in Finite Fuzzy Sets. *Information Sciences*, Elsevier North-Holland Inc., v.15, p. 159-168, 1978.

TURTLE, H.W. and CROFT, W.B. Uncertainty in Information retrieval Systems. In: UNCERTAINTY MANAGEMENT IN INFORMATION SYSTEMS - FROM NEEDS TO SOLUTIONS. USA: Kluwer Academic Publishers; Motro, A. and Smets P. (eds.); 1997. P.189-224.

VON ALTROCK, C. *Fuzzy Logic and NeuroFuzzy Applications Explained*. USA: Prentice Hall, 1995.

ZADEH, L. A. Fuzzy Sets. *Information and Control*, v.8, n.3, p. 338-353, June 1965.

ZADEH, L. A. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. *IEEE Transactions on Systems, Man and Cybernetics*, v.SMC-3, n.1, p.28-44, January 1973.