# A LÓGICA FUZZY APLICADA À RECUPERAÇÃO DE INFORMAÇÃO

relatos de experiência

*Edberto Ferneda* *
*Guilherme Ataíde Dias* **

**RESUMO**

O processo de recuperação de informação envolve conceitos subjetivos, imprecisos e vagos tais como "necessidade de informação", "relevância", além do próprio conceito de "informação". Os principais modelos de recuperação de informação tratam tais conceitos de maneira exata, representados por um único valor numérico. A lógica *fuzzy*, ao operar com a incerteza dos fenômenos da natureza de uma forma sistemática e rigorosa, representa uma alternativa promissora na solução de alguns problemas relacionados à recuperação de informação. Este trabalho apresenta a lógica *fuzzy* e alguns exemplos de sua utilização em sistemas de recuperação de informação (SRI).

**Palavras-chave:** Recuperação de Informação. Lógica Fuzzy. Conjuntos Fuzzy.

* Universidade Estadual Paulista – UNESP-Marília. Departamento de Ciência da Informação. E-mail: ferneda@marilia.unesp.bf
** Universidade Federal da Paraíba – UFPB. Departamento de Ciência da Informação. E-mail: Guilherme@dci.ccsa.ufpb.br

## 1 INTRODUÇÃO

As pesquisas em sistemas computacionais de recuperação de informação datam da década de 1950. Porém, com o surgimento da Web no início dos anos de 1990, a importância desses sistemas cresce continuamente à medida que cresce o número de documentos (páginas, imagens, sons, vídeos) prontamente compartilhados e disponíveis.

Em mais de meio século de pesquisas, aliado a um acelerado avanço das tecnologias de informação e comunicação (TIC), inúmeras ideias, conceitos e técnicas de recuperação de informação foram propostos e

desenvolvidos. Porém, a busca por informações relevantes e úteis ainda é uma tarefa bastante árdua.

Recuperar informação implica em operar seletivamente um estoque de informação, o que envolve processos cognitivos difíceis de serem formalizados. Assim, a utilização de recursos computacionais nessa tarefa parte de inevitáveis simplificações teóricas e de adequações de conceitos subjetivos tais como "relevância" e "necessidade de informação", além do próprio conceito de informação.

Em vista da imprecisão inerente ao processo de recuperação de informação, pesquisadores tentam abordar os problemas relacionados à essa área utilizando a lógica nebulosa, ou lógica fuzzy, a fim de representar com mais propriedade os conceitos envolvidos nesse processo.

Esse trabalho visa apresentar a lógica fuzzy como uma forma de melhor representar a imprecisão envolvida no processo de recuperação de informação. Inicialmente será apresentada a recuperação de informação como uma área de pesquisa multidisciplinar, mas com particular interesse da Ciência da Informação e da Ciência da Computação.

Serão apresentados os principais elementos envolvidos no processo de recuperação de informação e os principais modelos computacionais desenvolvidos entre as décadas 1960 e 1970, os modelos clássicos. Por fim, serão apresentados os conceitos de Lógica e Conjuntos Fuzzy para, em seguida, apresentar formas de utilização de tais conceitos na recuperação de informação.

## 2 RECUPERAÇÃO DE INFORMAÇÃO

Recuperar uma informação consiste em identificar, em um acervo documental, quais os documentos satisfazem total ou parcialmente a uma determinada necessidade de informação do usuário. Em princípio, considera-se que o usuário está interessado em recuperar informação sobre um determinado assunto e não documentos, embora seja nestes que a informação está registrada.

A Recuperação de Informação se estabeleceu como área de pesquisa em 1951, quando Calvin Mooers criou o termo (Information Retrieval) e definiu os problemas a serem abordados por esta nova disciplina. "[A Recuperação de Informação] trata dos aspectos

intelectuais da descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação (MOOERS, 1951, p. 21)".

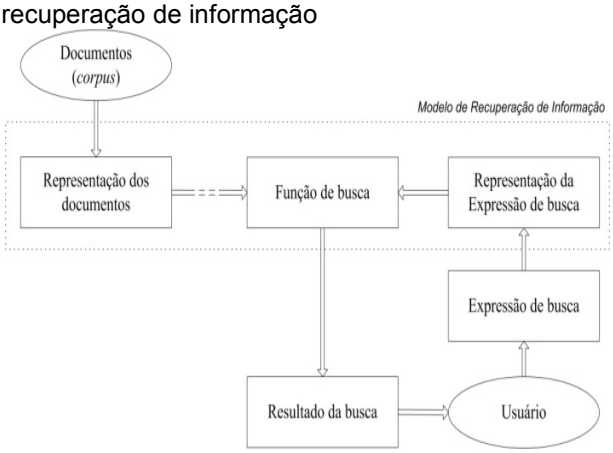
Para Saracevic (1999), a Recuperação de Informação pode ser considerada a vertente tecnológica da Ciência da Informação e é resultado da relação desta com a Ciência da Computação. Diversos processos criados e desenvolvidos ao longo da história da Ciência da Informação estão direta ou indiretamente relacionados à recuperação de informação.

Na Ciência da Computação a Recuperação de Informação se firmou como uma área de pesquisa autônoma, cujo interesse está centrado no desenvolvimento de ferramentas para o tratamento de fontes de informação não estruturadas ou semiestruturadas. É tema de interesse de uma imensa comunidade de pesquisadores de todas as partes do mundo e abriga uma grande quantidade de vertentes, abordagens e metodologias. Conta com periódicos e eventos direcionados especificamente para essa área.

## 2.1 O processo de recuperação de informação

O processo de recuperação de informação consiste em identificar, no conjunto de documentos (corpus) de um sistema, quais atendem à necessidade de informação do usuário. Uma representação simplificada do processo de recuperação de informação é apresentada na Figura 1.

**Figura 1 –** Representação do processo de recuperação de informação




Fonte: Adaptado de Ferneda (2012, p.14)

No ambiente digital que vem se configurando nas últimas décadas, o conceito de documento teve que se abstrair de seu suporte e centrar-se no que lhe é essencial: a informação. Nesse novo cenário, textos, imagens, sons, vídeos, páginas Web e diversos outros objetos digitais requerem diferentes tipos de tratamento e representação para uma

recuperação de informação eficaz (BURKE, 1999).

Buckland (1991) define o conceito de "informação como coisa" e argumenta que os acervos dos sistemas de informação são registros relacionados a coisas ou objetos. Nesses sistemas, informação está vinculada ao objeto que a contém. Por sua vez o termo documento, entendido como coisa informativa, incluiria, por exemplo, objetos, artefatos, imagens e sons.

Independente dos tipos de documentos gerenciados por um sistema de informação, a eficiência da recuperação é dependente da forma como esses documentos estão representados. A representação de um documento tem por objetivo identificar e descrever resumidamente o seu conteúdo informacional, ao mesmo tempo em que define seus pontos de acesso para a busca em um sistema de informação. A tarefa de representar os documentos é feita em um tempo anterior à execução de qualquer busca. No esquema da Figura 1, a existência de uma seta seccionada tenta mostrar essa assincronia.

Em um SRI usuário expressa sua necessidade de informação por meio de uma expressão de busca, composta geralmente por um conjunto de termos que representa linguisticamente a sua necessidade de informação. A principal dificuldade está em predizer os termos que foram usados para representar os documentos que satisfarão sua necessidade, e ao mesmo tempo evitar a recuperação de documentos não relevantes. Para isso é necessário que o usuário tenha um relativo conhecimento do vocabulário ou da terminologia do domínio de conhecimento de seu interesse.

Um SRI pode oferecer diversos recursos para facilitar o usuário na tarefa de expressar a sua necessidade de informação por meio da formulação de uma expressão de busca. Porém, para que seja possível uma comparação entre a expressão de busca e cada um dos documentos do corpus é necessário que as representações (da busca e dos documentos) sejam similares.

No centro do processo de recuperação de informação está a função de busca, que compara as representações dos documentos com a representação da expressão de busca e recupera os itens que supostamente fornecerão informações úteis ou relevantes para o usuário. O resultado de

uma busca é geralmente composto por uma lista de referências a documentos, ordenada pelo grau de relevância calculada pela função de busca.

Um modelo de recuperação de informação é a especificação formal de três elementos: a representação dos documentos, a representação da expressão de busca e a função de busca (FERNEDA, 2012, p.20). De maneira mais formal, Baeza-Yates e Ribeira-Neto (2011, p.58) definem modelo de recuperação de informação como uma quadrupla:

[**D**, **Q**, F, R(qi, dj)]

a) **D** é um conjunto composto por visões lógicas (representações) dos documentos no corpus;
b) **Q** é um conjunto composto de visões lógicas das necessidades de informação dos usuários;
c) F é um framework para a modelagem de representações dos documentos, consultas e seus relacionamentos;
d) $R(q_i, d_j)$ é uma função de ordenamento (ranking) que atribui um número real à relação entre uma representação da consulta $q_i$ de **Q** e a representação de um documento $d_j$ de **D**.

Apesar de alguns dos modelos de recuperação de informação terem sido criados entre as décadas de 1960 e 1970, as suas principais ideias ainda estão presentes na maioria dos sistemas de recuperação atuais e nos mecanismos de busca da Web.

## 2.2 Modelos de Recuperação de Informação

Os chamados "modelos clássicos" de recuperação de informação comportam propostas que serviram de base para o desenvolvimento de diversos outros modelos e algumas técnicas que até hoje são utilizadas. São eles: modelo booleano, modelo espaço vetorial e o modelo probabilístico.

### 2.2.1 Modelo Booleano

No modelo booleano um documento é representado por um conjunto de termos de indexação. As buscas são formuladas por meio de uma expressão booleana composta por termos ligados através dos operadores lógicos AND, OR e NOT, e apresentam como resultado o conjunto de documentos cuja representação satisfaz as restrições lógicas da expressão de busca.

Uma expressão conjuntiva de enunciado $\mathbf{t_1}$ **AND** $\mathbf{t_2}$ recuperará documentos indexados por ambos os termos ($\mathbf{t_1}$ e $\mathbf{t_2}$). Uma expressão disjuntiva $\mathbf{t_1}$ **OR** $\mathbf{t_2}$ recuperará o conjunto dos documentos indexados pelo termo $\mathbf{t_1}$ ou pelo termo $\mathbf{t_2}$. Uma expressão que utiliza apenas um termo $\mathbf{t_1}$ terá como resultado o conjunto de documentos indexados por esse termo. A expressão **NOT** $\mathbf{t_1}$ recuperará os documentos que não são indexados pelo termo $\mathbf{t_1}$. As expressões $\mathbf{t_1}$ **NOT** $\mathbf{t_2}$ ou $\mathbf{t_1}$ **AND NOT** $\mathbf{t_2}$ terão como resultado o conjunto dos documentos que são indexados por $\mathbf{t_1}$ e que não são indexados por $\mathbf{t_2}$.

Termos e operadores booleanos podem ser combinados para especificar buscas mais detalhadas ou restritivas. Como a ordem de execução das operações lógicas de uma expressão influencia no resultado da busca, muitas vezes é necessário explicitar essa ordem delimitando partes da expressão por meio de parênteses. A definição de expressões complexas exige um conhecimento profundo da lógica booleana. O conhecimento da lógica booleana é importante também para entender e avaliar os resultados obtidos em uma busca.

O modelo booleano possui diversas limitações. Algumas delas são:

a) Não existe uma forma de atribuir importância relativa (pesos) aos termos de indexação dos documentos nem aos diferentes termos da expressão de busca. Assume-se implicitamente que todos os termos possuem o mesmo peso.
b) O resultado de uma busca booleana se caracteriza por uma simples partição do corpus em dois subconjuntos: os documentos que atendem à expressão de busca e aqueles que não atendem. Presume-se que todos os documentos recuperados são de igual relevância, não havendo nenhum mecanismo pelo qual os documentos possam ser ordenados;
c) Sem um treinamento apropriado, o usuário leigo será capaz de formular somente buscas simples. Para buscas que exijam expressões mais complexas é necessário um conhecimento sólido da lógica booleana.

Apesar de suas limitações, muitos sistemas se desenvolveram utilizando o modelo booleano como ponto de partida para a criação de novos recursos de recuperação. Neste sentido o modelo booleano pode ser considerado o modelo mais utilizado nos sistemas de recuperação de informação e nos mecanismos de busca da Web.

### 2.2.2 Modelo Espaço Vetorial

No modelo espaço vetorial, ou apenas modelo vetorial, um documento é representado por um vetor onde cada elemento representa o peso, ou relevância, do respectivo termo de indexação na representação do conteúdo informacional do documento. Da mesma forma que os documentos, uma expressão de busca também é representada por um vetor numérico onde cada elemento representa a importância (peso) do respectivo na representação da necessidade de informação do usuário.

A utilização da mesma representação vetorial tanto para os documentos como para as expressões de busca permite calcular o grau de similaridade entre uma expressão de busca e cada um dos documentos do corpus, ou mesmo entre dois documentos, por meio do cálculo de distância entre dois vetores. Em um espaço vetorial contendo **N** dimensões, a similaridade (**sim**) entre um documento $\mathbf{d_j}$ e uma expressão de busca **q** pode ser calculada utilizando a seguinte fórmula (SALTON; McGILL, 1983, p.121):

$$sim(d_j, q) = \frac{\sum_{i=1}^{N} (w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{N} w^2_{i,j}} \times \sqrt{\sum_{i=1}^{N} w^2_{i,q}}}$$

onde $w_{i,j}$ é o peso do i-ésimo termo do documento $d_j$ e $w_{i,q}$ é o peso do i-ésimo termo da expressão de busca q.

Os valores da similaridade entre uma expressão de busca e cada um dos documentos do corpus são utilizados no ordenamento dos documentos resultantes. Portanto, no modelo vetorial o resultado de uma busca é um conjunto de documentos ordenados pelo grau de similaridade entre cada documento e a expressão de busca.

Um dos maiores méritos do modelo vetorial é a definição de um dos componentes essenciais de qualquer teoria científica: um modelo conceitual. Este modelo serviu como base para o desenvolvimento de uma teoria que alimentou uma grande quantidade de

pesquisas e resultou em sistemas como o SMART (SALTON, 1971).

### 2.2.3 Modelo Probabilístico

O modelo probabilístico foi proposto inicialmente por Maron e Kuhns (1960) e posteriormente explorado por diversos outros pesquisadores, tais como Robertson e Jones (1976). A ideia é tratar o processo de recuperação de informação como um processo probabilístico, já que é caracterizado por seu grau de incerteza no julgamento de relevância dos documentos em relação a uma expressão de busca. Assim, é mais realista pensar em uma probabilidade de relevância do que em uma pretensa relevância exata, como a utilizada nos modelos booleano e vetorial.

A partir de uma expressão de busca, composta por um ou mais termos, o usuário expressa sua necessidade de informação e a submete ao sistema. Por meio de cálculos de probabilidade o sistema calcula, para cada documento do corpus, um valor numérico (similaridade), que representa a provável relevância do documento para a consulta. Esse valor é utilizado para ordenar os resultados da busca. Tendo um primeiro conjunto de documentos, o usuário pode marca alguns deles que considera verdadeiramente relevantes para a sua necessidade. O conjunto de documentos marcados pode ser então submetido ao sistema, permitindo fornecer resultados mais precisos. Esse processo, denominado relevance feedback, pode ser repetido até que o usuário se sinta satisfeito com os resultados.

Uma virtude do modelo probabilístico está em reconhecer que a atribuição de relevância é uma tarefa do usuário, pois é o único modelo que incorpora explicitamente o processo de relevance feedback como base para a sua operacionalização.

Os três modelos clássicos compartilham a ideia de que a relevância de um documento pode ser definida de forma exata, representado por um valor numérico. Porém, o processo de recuperação de informação é inerentemente impreciso. A modelagem matemática desse processo só é possível somente através de simplificações teóricas e da adequação de conceitos tipicamente subjetivos como “necessidade de informação”, “relevância”, além do próprio conceito de “informação”. Esses e outros conceitos relacionados à

recuperação de informação são vagos e imprecisos e devem ser representados por meio de uma lógica que consiga capturar tais características.

## 3 LÓGICA FUZZY

A lógica aristotélica é uma forte presença na cultura ocidental e está profundamente enraizada em nossa forma de pensar. Uma determinada afirmação é verdadeira ou falsa; uma pessoa ou é amiga ou inimiga, feia ou bonita. Na ciência a verdade e a precisão estão intimamente ligadas e são partes indispensáveis do método científico. Se algo não é absolutamente correto então não é verdade. Porém, observa-se um considerável descompasso entre a realidade e essa visão bivalente do mundo. O mundo real contém uma infinidade de gradações entre o preto e o branco, entre o feio e o bonito, entre o verdadeiro e o falso. O mundo real é multivalente e analógico. Verdade e precisão absolutas existem apenas em casos extremos.

A própria comunicação humana é inerentemente vaga e imprecisa. Quando se diz que uma determinada pessoa é alta, o que se está querendo dizer precisamente: 1,70m 1,80m ou 1,90m? A afirmação de que uma determinada pessoa é bonita ou feita põe em jogo um grande conjunto de variáveis difíceis de equacionar. Quando os seres humanos pensam em altura ou em beleza eles normalmente não têm uma medida em mente, mas uma definição nebulosa, vaga.

O objetivo da lógica fuzzy é capturar e operar com a diversidade, a incerteza e as verdades parciais dos fenômenos da natureza de uma forma sistemática e rigorosa (SHAW; SIMÕES, 1999).

### 3.1 Conjuntos fuzzy

Zadeh (1965) propôs uma nova teoria de conjuntos na qual não existem descontinuidades, ou seja, não há uma distinção abrupta entre elementos pertencentes e não pertencentes a um determinado conjunto. Essa nova teoria é derivada da lógica fuzzy: os Conjuntos Nebulosos (Fuzzy Sets).

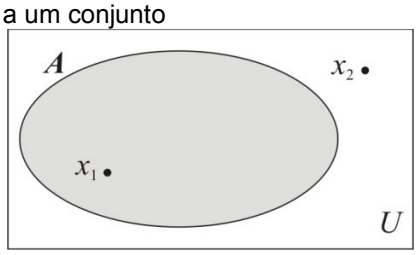
Na teoria matemática dos conjuntos, para indicar que um elemento **x** pertence a um conjunto **A**, utiliza-se a expressão $\mathbf{x} \in \mathbf{A}$. Poderia-se também utilizar a função $\mu_A(x)$, cujo valor indica se

o elemento **x** pertence ou não ao conjunto **A** . Neste caso $\mu_A(x)$ é uma função bivalente que somente resulta 1 (um) ou zero, dependendo se o elemento **x** pertence ou não ao conjunto **A**:

$$\mu_A(x) = \begin{cases} 1 \text{ se } x \in A \\ 0 \text{ se } x \notin A \end{cases}$$

Na Figura 2 observa-se que, se o elemento $\mathbf{x_2}$ for movido em direção ao elemento $\mathbf{x_1}$, no limite do conjunto **A** ocorrerá subitamente uma alteração de seu estado, passando de não-membro para membro do conjunto **A**.

**Figura 2 -** Pertinência de um elemento em relação a um conjunto

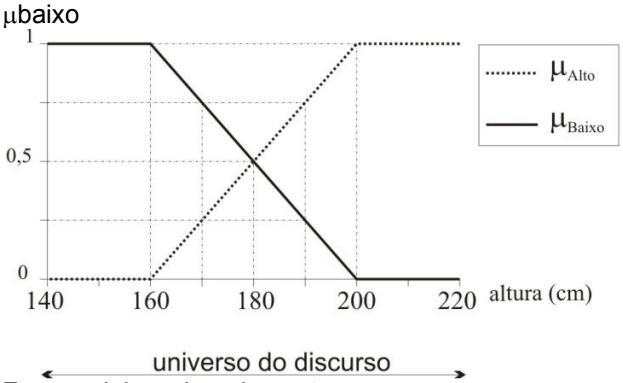



Fonte: Elaborado pelos autores

Na lógica fuzzy um elemento pode ser membro de um conjunto apenas parcialmente. Um valor entre zero e um (1) indicará o quanto o elemento é membro do conjunto.

A teoria dos conjuntos fuzzy é baseada no fato de que os conjuntos existentes no mundo real não possuem limites precisos. Um conjunto fuzzy é um agrupamento indefinido de elementos no qual a transição de cada elemento de não-membro para membro do conjunto é gradual. Esse grau de imprecisão de um elemento pode ser visto como uma medida de possibilidade, ou seja, a possibilidade de que um elemento seja membro do conjunto.

**Figura 3 -** Representação das funções μalto e μbaixo




**Fonte:** elaborado pelos autores

No exemplo da Figura 3 o conjunto dos diversos valores das alturas de uma pessoa é denominado universo do discurso. Todo conjunto fuzzy é na realidade um subconjunto do universo do discurso. Um subconjunto A do universo do discurso U é caracterizado por uma função $\mu_A$ que associa a cada elemento x de U um número $\mu_A(x)$ entre 0 e 1. Assim, temos:

$$A = \{x, \mu_A(x)\} \mid x \in U$$

onde $\mu_A(x)$ resulta um valor numérico entre zero e um que representa o quando o elemento x pertence ao conjunto A.

Supondo que A seja o conjunto de pessoas altas e $\mathbf{x_1}$ e $\mathbf{x_2}$ representam duas pessoas com 190 cm e 170 cm de altura, respectivamente. O subconjunto A será caracterizado pela função $\mu_A(x)$, que associa a cada elemento $x_1$ e $x_2$ do universo do discurso um número, respectivamente $\mu_A(x_1)$ e $\mu_A(x_2)$. No gráfico da Figura 2 teremos $\mu_A(x_1)$ igual a 0,75 (75%) e $\mu_A(x_2)$ igual a 0,25 (25%). Portanto, no exemplo, uma pessoa com 190cm é 75% alta e 25% baixa. Uma pessoa com 170cm é apenas 25% alta e 75% baixa. Ou seja, em um conjunto fuzzy um mesmo objeto pode pertencer a dois ou mais conjuntos com diferentes graus. Uma pessoa que mede 180 cm é simultaneamente 50% alta e 50% baixa ($\mu_{alta}(180)=\mu_{baixa}(180)=0.5$).

As operações mais utilizadas nos conjuntos fuzzy são: complemento, união e interseção e são definidas como segue:

Complemento: $\mu_{\bar{A}}(u) = 1 - \mu_A(u)$

União: $\mu_{A\cup B}(u) = \max(\mu_A(u), \mu_B$

Inserseção: $\mu_{A\cap B}(u) = \min(\mu_A(u), \mu_B$

Utilizando a Figura 2, essas operações são exemplificadas abaixo:

$$\mu_{\overline{baixo}}(170) = 1 - \mu_{baixo}(170) = 1 - 0.75 = \mathbf{0.25}$$

$$\mu_{baixo\cup alto}(170) = \max(\mu_{baixo}(170), \mu_{alto}(170)) = \max(0.75, 0,25) = \mathbf{0.75}$$

$$\mu_{baixo\cap alto}(170) = \min(\mu_{baixo}(170), \mu_{alto}(170)) = \min(0.75, 0.25) = \mathbf{0.25}$$

A teoria fuzzy possibilita a definição de classes de elementos em situações onde não é possível uma delimitação precisa e natural de suas fronteiras. Este ambiente teórico é capaz de representar de forma mais eficiente a imprecisão das entidades envolvidas em um sistema de recuperação de informação.

## 4 A LÓGICA FUZZY NA RECUPERAÇÃO DE INFORMAÇÃO

Ao utilizarmos um sistema de recuperação de informação por vezes não temos uma ideia clara, exata, da informação que precisamos, nem tampouco temos conhecimento dos termos que devemos utilizar para traduzir nossa necessidade de informação em uma expressão de busca que resulte em um conjunto de documentos relevantes. O próprio conceito de relevância é vago e impreciso, além de ser pessoal. Assim, um mesmo documento pode variar de absolutamente relevante à totalmente

irrelevante, dependendo do usuário do sistema (BORDOGNA; PASI, 2000).

Um exemplo de aplicação dos conjuntos fuzzy na recuperação de informação é dado por Bordogna e Pasi (1995). Esses autores aplicam a lógica fuzzy da descrição dos documentos de um sistema de recuperação de informação.

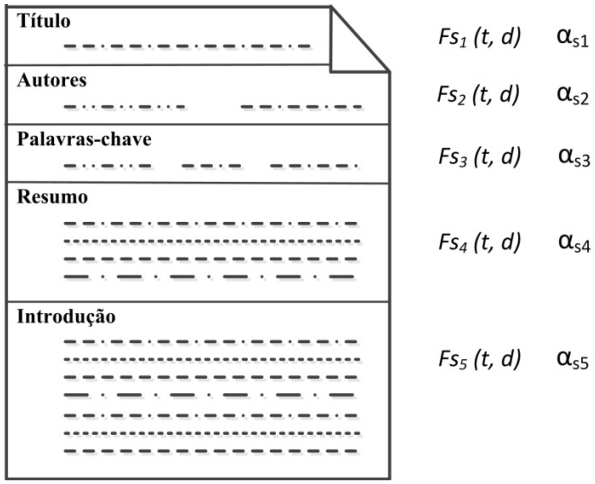
Um documento pode ser visto como um conjunto fuzzy de termos, { $\mu$(t)/t }, cujos pesos dependem do documento e do termo em questão, isto é: $\mu$(t)=F(d,t). Portanto, a representação fuzzy de um documento é baseada na definição de uma função F(d, t) que produz um valor numérico que representa o peso do termo t para o documento d.

O peso associado a um termo expressa o quanto esse termo é significativo na descrição do conteúdo do documento. A qualidade da recuperação depende em grande parte da função adotada para calcular os pesos dos termos de indexação (SALTON; BUCKLEY, 1988). Geralmente esta função baseia-se no cálculo da freqüência de ocorrência dos termos em todo o texto, e fornece uma representação estática do documento. O cálculo dos pesos não considera que em muitos casos os documentos podem estar estruturados em sub-partes lógicas ou seções, e que as ocorrências de um termo podem assumir significados diferentes dependendo da seção onde ele aparece. Um artigo científico, por exemplo, geralmente está organizado em título, autores, palavras-chave, resumo, referências, etc. Uma única ocorrência de um termo no título sugere que o artigo discorre sobre o conceito expresso pelo termo. As seções de um documento podem assumir diferentes graus de importância dependendo da necessidade do usuário. Quando, por exemplo, o usuário está procurando artigos escritos por uma determinada pessoa, a parte mais importante a ser analisada é a seção de autores. Quando se procura artigos de um determinado assunto, o título, as palavras-chaves, o resumo e a introdução assumem maior importância.

Bordogna e Pasi (1995) propõem uma representação fuzzy para documentos estruturados que pode ser ajustada de acordo com os interesses do usuário. A importância de um termo t em um documento d é calculada pela avaliação da importância de t em cada uma das seções de d. Isto é feito através

da aplicação de uma função $F_{Si}(d, t)$ que expressa o grau de pertinência do termo t na seção $S_i$ do documento d, como ilustrado na Figura 4.

**Figura 4 -** Representação fuzzy de um documento estruturado




**Fonte:** adaptado (BORDOGNA; PASI, 1995)

Para cada seção $S_i$ o usuário pode associar uma importância numérica $\alpha_{Si}$ que será usada para enfatizar a função $F_{Si}(t,d)$. Para se obter um grau de pertinência ou relevância de um determinado termo em relação a um documento, os graus de pertinência do termo em cada uma das seções $F_{S1}(d,t)$, $F_{S2}(d,t)$,...$F_{Sn}(d,t)$ são agregados por meio de uma função, que pode ser selecionada pelo usuário entre um conjunto pré-definido de "quantificadores lingüísticos" tais como all, least one, at least about k, all (YAGER, 1988). O quantificador lingüístico indica o número de seções em que um termo deva aparecer para que o documento seja considerado relevante. Esta representação fuzzy de documentos foi implementada em um sistema denominado DOMINO (BORDOGNA et al, 1990) e mostrou ser mais eficaz em relação a outros tipos de representação fuzzy.

Utilizando idéia semelhante, Molinari e Pasi (1996) propõem um método de indexação de documentos HTML baseado na estrutura sintática dessa linguagem de marcação. Para cada seção de um documento HTML, delimitada pelas marcações (tags), é associado um grau de importância. Pode-se supor, por exemplo, que quanto maior o tamanho dos caracteres de um trecho do texto maior a importância atribuída a esse trecho. Da mesma forma, uma palavra em negrito ou itálico geralmente representa um destaque dado pelo autor da página HTML para uma palavra. Assim, para cada tag pode ser associado um valor numérico que expressa a sua importância para o documento. O peso de um determinado termo em relação a um determinado documento é obtido através de uma função de agregação que considera a importância de cada tag do documento onde o termo aparece.

## 5 CONSIDERAÇÕES FINAIS

A lógica fuzzy objetiva capturar e operar com a imprecisão e a incerteza dos fenômenos da natureza de uma forma sistemática e rigorosa. Ao considerar a subjetividade e imprecisão, a lógica fuzzy permite representar mais adequadamente o processo de recuperação de informação.

A utilização da lógica fuzzy na recuperação de informação é discutida principalmente em comunidades dedicadas às pesquisas sobre teoria fuzzy, não tendo, na maior parte das vezes, ligação com os pesquisadores da área de recuperação de informação. Apesar disso, a aplicação da lógica fuzzy na recuperação de informação traz novos conceitos e ideias promissoras que podem resultar avanços significativos nessa área.

Espera-se que a partir do desenvolvimento de novos e mais poderosos recursos computacionais, novas abordagens, métodos e modelos possam ser efetivamente aplicados aos problemas da recuperação de informação. Nesse contexto, a lógica fuzzy se apresenta como uma alternativa promissora para futuras pesquisas.

## *FUZZY LOGIC APPLIED TO INFORMATION RETRIEVAL*

***ABSTRACT***

*The information retrieval process involves subjective, imprecise and vague concepts, such as "information need", "relevance", and the very concept of "information". The main information retrieval models treat these concepts accurately, represented by a single numerical value. The fuzzy logic, while operating with the uncertainty of natural phenomena in a systematic and rigorous manner, represents a promising alternative to solve some problems related to information retrieval. This paper presents the fuzzy logic and some examples of its use in information retrieval systems (IRS).*

***Keywords:*** *Information Retrieval. Fuzzy Logic. Fuzzy Sets.*

## REFERÊNCIAS


BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**, 2.ed. [S.l.]: Addison-Wesley, 2011.

BORDOGNA, G.; PASI, G. Modeling vagueness in information retrieval. In: CRESTANI, F.; AGOSTI, M.; PASI, G. (eds) **Lectures on Information Retrieva**l. [S.l.]: Springer-Verlag, 2000.

BORDOGNA, G.; PASI, G. Controlling Information Retrieval through a user adaptive representation of documents. **International Journal of Approximate Reasoning**, v.12, 1995.

BORDOGNA, G. et al. A system architecture for multimedia information retrieval. **Journal of Information Science**. v. 16, n. 2, 1990.

BUCKLAND, M. K. Information as thing. **Journal of the American Society of Information Science**, v.42, n.5, 1991.

BURKE, M. A. **Organization of multimedia resources: principle and practice of information retrieval**. Aldershot: Gower, 1999.

FERNEDA, E. **Introdução aos Modelos Computacionais de Recuperação de Informação**. Rio de Janeiro: Ciência Moderna, 2012.

MOLINARI, A. ; PASI, G. A Fuzzy Representation of HTML Documents for Information Retrieval Systems. IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS. New Orleans, 1996. **Proceedings…** New Orleans: [S.n.], 1996.

MOOERS, C. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, v. 2, n. 1, 1951.

SALTON, G.; McGILL, J. M. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill, 1983.

SALTON, G.; BUCKLEY, C. Term-Weighting Approaches in Automatic Text Retrieval. **Information Processing and Management**, v. 24, n. 5, 1988.

SARACEVIC, T. Information Science. **Journal of the American Society for Information Science**, v. 50, n. 12, 1999.

SHAW, I. S.; SIMÕES, M. G. **Controle e modelagem fuzzy**. São Paulo: Edgard Blücher, 1999.

YAGER, R.R. On ordered weighted averaging aggregation operators in multi-criteria decision making, **IEEE transactions on Systems, Man and Cybernetics**, v. 18, 1988.

ZADEH, L.A. Fuzzy sets. **Information and Control**, v. 8, n. 3, 1965