

# ESTUDO COMPARATIVO DE JULGAMENTOS DE RELEVÂNCIA DO USUÁRIO E NÃO-USUÁRIO DE SERVIÇOS DE D. S. I.

Regina Célia Figueiredo  
Divisão de Informação e Documentação Científicas  
Instituto de Energia Atômica de São Paulo

## 1 INTRODUÇÃO

Em agosto de 1974, foi implantado, na Divisão de Informação e Documentação Científicas (DIDC) do Instituto de Energia Atômica (IEA), de São Paulo, um serviço manual de disseminação seletiva da informação (SDI).

O serviço tem por objetivo a recuperação e disseminação periódica e seletiva de informações para atender às pesquisas prioritárias sendo desenvolvidas pela comunidade científica do IEA. Atende às mais diversas áreas de assunto relacionadas à energia nuclear: física, química, metalurgia, radiobiologia, proteção radiológica, engenharia de reatores, processamento de material radioativo, etc.

Na área nuclear, existe um sistema de informação internacional, International Nuclear Information System — INIS —, patrocinado pela Agência Internacional de Energia Atômica (IAEA). Os países membros da IAEA, participantes do INIS, coletam e promovem a entrada de dados ("input") referentes à literatura nuclear produzida em cada país. A IAEA processa essa informação e a divulga sob as formas de bibliografia impressa — o INIS Atomindex — e fitas magnéticas. As fitas magnéticas são acessíveis aos países membros que desejam recuperar e disseminar as informações contidas nessas fitas, em bases nacionais.

\* *Extrato da dissertação apresentada ao Instituto Brasileiro de Informação em Ciência e Tecnologia para obtenção do grau de mestre em Ciência da Informação.*

## RESUMO

*Os julgamentos de relevância, atribuídos a um grupo de documentos da área de energia nuclear, por usuários de um serviço de disseminação seletiva da informação foram comparados àqueles atribuídos por um não-usuário. A avaliação de relevância foi feita pelos títulos dos documentos sendo observada a utilidade dos mesmos como determinantes da relevância ou não dos documentos. Os julgamentos dos usuários e do não-usuário foram similares, sendo aceitável a porcentagem de discordâncias. A porcentagem de concordâncias não foi maior devido à diferença de pontos de vista entre usuário e não-usuário: o primeiro avalia o documento para atender a sua necessidade de informação e o segundo em função da questão proposta. A frequência de descritores do perfil nos títulos dos documentos não discrimina, por si só, os documentos relevantes dos não relevantes.*

*Descritores: Relevância; Usuário; Disseminação Seletiva da Informação.*

No Brasil, a Comissão Nacional de Energia Nuclear (CNEN), através de seu Centro de Informações Nucleares (CIN), desenvolveu um serviço de SDI automatizado, utilizando as fitas magnéticas do INIS. As fitas são recebidas e processadas quinzenalmente e a informação recuperada é enviada periodicamente aos usuários do sistema.

O serviço de SDI manual implantado em São Paulo abrange as mesmas áreas de interesse atendidas pelo SDI do CIN, sendo dirigido, entretanto, somente à comunidade científica do IEA.

A bibliotecária encarregada do SDI manual do IEA teve oportunidade, em determinada época, de participar das atividades desenvolvidas pelo serviço de SDI automatizado do CIN.

Em consequência da identidade de propósitos dos dois serviços, os perfis anteriormente atendidos pelo SDI manual do IEA foram incluídos no SDI automatizado do CIN.

Essa inclusão motivou a idéia de analisar julgamentos de relevância.

Com esse objetivo em mente, a bibliotecária comparou, durante um período de aproximadamente seis meses, os julgamentos de relevância dos usuários com os seus próprios (não-usuário), atribuídos a um mesmo grupo de documentos.

## 2 HIPÓTESES

Foram formuladas as seguintes hipóteses para o estudo:

- 1) Não há diferença entre os julgamentos de relevância emitidos pelos usuários e aqueles emitidos por um não-usuário, desde que este último tenha conhecimentos básicos dos assuntos cobertos pelo SDI e familiaridade com o enfoque das questões propostas.
- 2) São relevantes documentos cujos títulos contêm mais de dois descritores mencionados no perfil e, não-relevantes aqueles cujos títulos não contêm nenhum dos descritores do perfil.

## 3 DEFINIÇÃO DOS CONCEITOS

No contexto deste trabalho, entendeu-se por *relevância* a relação entre os documentos e a questão proposta pelo usuário. E por pertinência, a relação entre os documentos e a necessidade de informação do usuário.

Os *julgamentos de relevância* são as avaliações dos documentos recuperados sob o ponto de vista do interesse ou não-interesse dos documentos para a questão proposta.

## 4 REVISÃO DA LITERATURA

A revisão da literatura limitou-se a artigos que apresentam aspectos e fatores que influenciam julgamentos de relevância. Algumas experiências sobre comparações entre julgamentos de relevância do usuário e do sistema (ou não-usuário) e estudos sobre fatores que podem determinar os julgamentos de relevância do usuário foram particularmente revisados por estarem diretamente relacionados ao âmbito deste estudo. Devido à dificuldade de medir os julgamentos do usuário, a maior parte dos artigos lidam com medidas de relevância sob o ponto de vista do sistema e tentam aproximá-las a uma *medida mais restrita*.

A literatura referente aos conceitos de relevância e pertinência também foi analisada.

### 4.1 Relevância e Pertinência

BRADFORD<sup>(7)</sup> foi o primeiro a utilizar o termo "relevante", no contexto em que ele é usado hoje, quando falou em artigos "relevantes a um assunto".

Nas décadas de 40 e 50, MOOERS<sup>(12)</sup>, TAUBE<sup>(23)</sup> e PERRY<sup>(14)</sup> foram os primeiros a falar sobre relevância em relação à recuperação da informação. Eles se preocuparam mais com a não-relevância, que foi apontada como causa da má-utilização ou adequação dos sistemas.

Em 1958, durante a "International Conference for Scientific Information" (ICSI)<sup>(13)</sup>, o conceito de relevância foi discutido e explorado por diversos autores.

Várias definições de relevância aparecem na década de 60, principalmente como resultado da afirmativa de que não era muito claro o significado do termo, como afirmou SARACEVIC<sup>(20)</sup>. Muitas definições eram somente paráfrases de definições anteriores, numa tentativa de explicar as relações entre usuário/documento ou sistema/documentos.

Vários termos foram propostos por diversos autores para identificar diferentes aspectos de relevância. Pertinência, relevância do sistema, relevância do usuário etc. foram termos muitas vezes conceituados com o sentido de relevância.

Ao especificar os fatores, elementos e relações **que** a noção de relevância compreende, os cientistas da informação desenvolveram muitos pontos de vista **sobre** relevância, de acordo com as características que eram mais importantes na época em que foram **propostos**. Assim, de acordo com SARACEVIC<sup>(20)</sup>, pode-se ter a relevância do ponto de vista do sistema, do destino da informação, da literatura sobre o assunto, do conhecimento do assunto, e por uma visão pragmática. Esses pontos de vista não são mutuamente exclusivos; cada um deles completa alguns aspectos que **não** haviam sido reconhecidos anteriormente.

POLUSHKIN<sup>(17)</sup> afirmou que relevância é usualmente entendida como a característica do grau de correlação entre o conteúdo do documento recuperado e o conteúdo da questão. E definiu pertinência como a característica do grau de correlação entre o conteúdo do documento recuperado e o pedido de informação, em *conjunto com as características puramente subjetivas* do usuário específico.

SARACEVIC<sup>(20)</sup> definiu relevância como uma medida de eficácia do contato entre uma fonte e um destinatário num processo de comunicação. Salientou que o conceito de necessidade de informação deu origem à noção de pertinência.

FOSKETT<sup>(4)</sup> sugeriu que a *verdadeira distinção* é que relevância poderia ser tomada no sentido de pertencente ao campo/assunto/universo delimitado pelos termos da questão, tal como estabelecido pela opinião geral dos pesquisadores naquele campo; enquanto pertinência poderia ser definida como a edição de novas informações às já existentes na mente do usuário, as quais são úteis para ele no trabalho que o levou a formular a questão.

LANCASTER<sup>(11)</sup> sugeriu também uma distinção entre relevância e pertinência. Ele *distingue pertinência* como a relação entre um documento e uma necessidade de informação e *relevância* como a relação entre o documento e a *questão*.

HARMON<sup>(5)</sup> distinguiu relevância do usuário de relevância do sistema. Relevância do usuário pode ser vista como uma avaliação feita pelo usuário da extensão da relação entre a resposta do sistema e a sua necessidade de informação. A relevância do

sistema pode ser vista como a avaliação feita pelo sistema do grau de relação entre a informação contida no sistema e a questão proposta pelo usuário.

WILSON<sup>(24)</sup> tentou explicar relevância não como uma única noção, mas muitas. Definiu o que chamou de "relevância situacional" como sendo relevância para uma situação individual específica, para a situação como o indivíduo a vê, não como outros a vêem ou como ela "realmente" é. É também relevância em relação ao cabedal de conhecimentos de um indivíduo em particular e muda na medida em que esse cabedal se altera.

CUADRA e KATTER<sup>(25)</sup> salientam também o aspecto "situacional" da relevância. Afirmam os autores que os julgamentos de relevância têm sido tratados como se proviessem de uma "caixa preta", cujo conteúdo não tivesse interesse a não ser enquanto a caixa preta desempenhasse sua função, isto é, fornecesse um número de pontos que pudesse ser usado como um critério de medida. Para os autores, relevância é mais que uma simples contagem de pontos e pode ser influenciada por um sem número de variáveis.

#### 4.2 Experimentos com Julgamentos de Relevância

Os experimentos sobre relevância foram desenvolvidos de acordo com cinco classes gerais de variáveis, como foi revisado por SARACEVIC<sup>(26)</sup>: 1) documentos e representações de documentos; 2) questões ("queries"); 3) situações e condições de julgamento; 4) formas de expressão; e 5) características humanas (pessoas).

Os autores dos estudos revisados neste trabalho fixaram sua atenção principalmente em documentos e suas representações e na consistência de julgamentos de relevância, assim como na concordância desses julgamentos na medida em que são afetados por certas características humanas.

GIFFORD e BAUMANIS<sup>(27)</sup> descrevem uma investigação empírica sobre o papel dos documentos nos julgamentos de relevância. O estudo foi baseado na pergunta: "O que em um documento provoca o julgamento do usuário?" A questão foi diretamente explorada, principalmente pela determinação de diferenças textuais significativas que pudessem ser encontradas entre documentos relevantes e não-relevantes.

Na amostra examinada pelos autores os documentos relevantes distinguiram-se com regularidade dos não-relevantes. Contudo, em relação à noção geral de relevância, ou melhor, ao que torna um documento relevante, nenhuma evidência dos atributos gerais foi encontrada. Os documentos relevantes para uma questão específica geralmente têm propriedades significativas em comum, mas isto não se estende de uma questão à outra.

TAGUE<sup>(28)</sup> tentou classificar as relações entre as palavras da questão e as palavras usadas para indexar documentos (respostas) relevantes, relativamente relevantes e não-relevantes.

O autor concluiu que: a) As palavras do título são mais efetivas como termos de avaliação; b) As palavras da questão são mais úteis para localizar respostas relevantes; c) A frequência de palavras da questão é maior em documentos relevantes que nos parcialmente relevantes ou não-relevantes; d) O número de palavras, mais do que a relação entre as palavras da questão original, é que caracteriza as respostas relevantes.

SARACEVIC, em outro estudo<sup>(29)</sup>, examinou aspectos relacionados com testes de sistemas de recuperação, entre eles: usuários e proposições de questões, e julgamentos de relevância baseados em diferentes formatos de saída. As medidas de avaliação revelaram que há maior concordância sobre o que não é relevante do que sobre o que é relevante.

O autor concluiu que os sistemas não atribuem relevância melhor do que um usuário mas que devem ter um desempenho tão bom quanto o do usuário. Afirmou ainda que diferentes formatos têm um efeito significativo sobre julgamentos de relevância do usuário.

RATH et alii<sup>(30)</sup> realizaram um estudo comparativo entre quatro tipos de indicadores léxicos de conteúdo (títulos, dois tipos de resumos e textos). Concluíram os autores que o uso de títulos sem resumo adicional leva a julgamentos de relevância discordantes. Não houve diferença significativa entre avaliações usando textos completos e resumos.

O'CONNOR<sup>(31, 32)</sup> conduziu diversos estudos para explicar discordância nos julgamentos de relevância e os fatores que estavam relacionados e/ou causando essas discordâncias. O autor decidiu que as causas básicas de discordância sobre relevância são as diferenças de interpretação das requisições ou documentos, e não fatores tais como educação, experiência, função etc. dos julgadores, e o que eles consideram como o objetivo, ambiente e momento da proposição da questão.

BARHYDT<sup>(33)</sup> realizou um estudo para testar a eficácia de julgamentos de relevância de não-usuários sobre respostas e questões apresentadas a um centro piloto. Concluiu que o usuário é o melhor árbitro da relevância do documento e que o especialista do assunto não é melhor (ou pior) que o especialista do sistema ao avaliar documentos.

RESNICK e SAVAGE<sup>(34)</sup> desenvolveram um experimento controlado para comparar a relativa consistência humana inter- e intra-assunto, com base em citações, resumos, termos de indexação e textos completos. Sob as condições de experimento os julgamentos foram consistentes. Restou como um problema, porém, a quantificação dos resultados dos julgamentos humanos.

#### 4.3 Métodos e Medidas

Os autores empregaram vários métodos e medidas para conceituar julgamentos de relevância e para determinar conexões semânticas entre questões e textos relevantes.

Como mencionou SARACEVIC<sup>(26)</sup>, as primeiras unidades quantitativas de medida foram: revocação ("recall") e precisão ("precisión"), definidas por KENT et alii<sup>(27)</sup>. Revocação é a razão das respostas relevantes recuperadas sobre o total de respostas relevantes do sistema. Precisão é a razão das respostas relevantes recuperadas sobre o total de respostas recuperadas.

BARHYDT<sup>(1)</sup> e SARACEVIC<sup>(21)</sup> apresentaram algumas medidas de avaliação do desempenho de sistemas: sensibilidade ("sensitivity"), especificidade ("specificity") e efetividade ("effectiveness"). A sensibilidade (Se) é a razão entre o número de documentos recuperados e o total de documentos relevantes no arquivo. Especificidade (Sp) é a razão entre o número de documentos não-relevantes não-recuperados e o total de documentos não-relevantes no arquivo. A efetividade (Es) é uma adição de Se e Sp menos uma constante 1.

GEBHARDT<sup>(4)</sup> propôs um modelo probabilístico para a avaliação de relevância dos documentos, considerando relevância como uma variável aleatória. Não apresentou, entretanto, uma escala padrão para a atribuição de julgamentos de relevância a documentos.

KING<sup>(10)</sup>, em uma revisão das medidas de desempenho dos sistemas, sugeriu uma representação da relação entre julgamentos de relevância do sistema e do usuário, na qual os valores de julgamento do usuário foram colocados num gráfico contra os valores correspondentes de relevância do sistema para cada documento de uma determinada questão.

GIFFORD e BAUMANIS<sup>(6)</sup> tentaram determinar diferenças textuais entre documentos relevantes e não-relevantes, identificando ocorrências comuns de palavras, frases e outras propriedades textuais nos documentos de cada questão. Uma declaração da combinação das propriedades foi demonstrada pela tabulação dessas propriedades para cada questão.

IVANKIN<sup>(8)</sup> analisou as conexões semânticas entre textos das proposições de questões e documentos relevantes, identificando as palavras-chaves comuns. O autor fez uma análise estatística das relações entre palavras-chave, suas tendências e as combinações entre todos os conjuntos de palavras-chave.

Para medir a consistência dos julgamentos de relevância humanos, RESNICK e SAVAGE<sup>(13)</sup> selecionaram um grupo de técnicos para avaliar um certo grupo de documentos. A avaliação foi repetida um mês depois e os resultados comparados.

HARMON<sup>(7)</sup> propôs um modelo de necessidade de informação do usuário que pode ser útil para interpretar a relevância do usuário. O modelo envolve conjuntos cognitivos dos usuários, que são organizados e ordenados de acordo com critérios específicos.

Segundo HARMON, a pesquisa deve ser vista como um processo cíclico ou modular envolvendo tanto a aquisição como a ordenação da informação. Mediante esse processo, a mudança da configuração da necessidade da informação influencia a natureza das perguntas

geradas. Por sua vez, as respostas às perguntas modificam a necessidade de informação.

Como se verificou através dessa revisão, a relevância ainda está longe de ser bem compreendida e conceituada. Cada um dos experimentos realizados dá uma nova dimensão para *um* enfoque das propriedades gerais, efeitos e fatores que afetam relevância.

Não há muito o que comparar entre os experimentos descritos na literatura, pois a população analisada, a coleta de dados, os métodos etc. variam bastante de um experimento para outro. Há, contudo, um objetivo comum: esclarecer relevância como uma medida de qualidade da informação recuperada e interpretar fatores que afetam julgamentos de relevância.

## 5 METODOLOGIA

### 5.1 Operacionalização dos Conceitos

Foram usadas duas medidas neste estudo: sensibilidade e especificidade (já definidas em 4. 3).

Quando os perfis que serviriam para o estudo estavam sendo trabalhados pelo SDI manual, a avaliação dos documentos a serem recuperados era feita pela bibliotecária, comparando-se o documento original com os descritores escolhidos pelos usuários. Eram recuperados os documentos que tratassem do assunto específico a ser pesquisado, ou seja, que contivessem alguns ou todos os descritores selecionados para definir e delimitar o assunto, de acordo com as relações entre os descritores mencionadas na descrição do assunto.

A informação recuperada era enviada ao usuário em forma de fichas contendo as referências bibliográficas. A cada recebimento de referências, o usuário analisava o conteúdo da informação recebida e respondia numa escala de dois pontos: de interesse ou não-interesse.

No caso do SDI automatizado, o procedimento é semelhante: o computador seleciona documentos de interesse para os perfis comparando os descritores selecionados no perfil com os descritores atribuídos ao documento na indexação. Na ocasião do estabelecimento do perfil, o usuário deve dar a cada descritor um peso numérico que indica o grau de interesse da palavra para sua pesquisa. O computador, ao comparar os descritores do documento com os do perfil, soma os pesos das palavras encontradas através de uma fórmula específica. Se a soma for maior ou igual a um patamar pré-estabelecido o documento é recuperado.

As referências selecionadas são enviadas aos usuários em forma de cartões. A cada bloco de cartões corresponde um cartão de interesse, onde o usuário deve registrar seus julgamentos de relevância para cada documento: "muito interesse" (M), "relativo interesse" (R) e "nenhum interesse" (N).

Neste estudo, decidiu-se comparar os julgamentos dos usuários com o do não-usuário (bibliotecária), atribuídos aos mesmos documentos.

Com os dados coletados pensava-se obter os seguintes indicadores:

- a) Qual o grau de relevância alcançado, considerando-se o total de documentos recuperados pelo SDI automatizado em dois níveis: as referências *M*, isoladamente, e, em outro grupo, as referências *M* e *R* em conjunto;
- b) Qual o índice de concordâncias e discordâncias entre as avaliações do sistema, do usuário ou do não-usuário.

### 5.2 Coleta de Dados

A coleta de dados foi feita durante um período de aproximadamente 6 meses, cobrindo os envios do SDI referentes aos fascículos 18 a 24 do volume 6, e 1 a 4 do volume 7, do INIS Atomindex.

A cada envio, quinzenal de SDI para os perfis em estudo, a bibliotecária avaliava os documentos recuperados pelo sistema, fazendo uma simulação do cartão-interesse, que seria devolvido posteriormente pelos usuários, com seus julgamentos de relevância. Esses dados eram comparados aos dos usuários.

A avaliação foi feita pelo título das referências.

Para a análise dos dados foi escolhida uma amostra dos documentos onde houve concordância de julgamentos e os documentos que causaram divergências extremas (*M* e *N*).

Os documentos avaliados em *R* não foram considerados para o estudo, pois não apresentaram uma característica significativa, isto é, o índice *R* apresentou-se em número quase idêntico tanto em relação ao índice *M* como ao índice *N*, sem mostrar uma tendência a se associar a um ou outro índice.

Para os documentos divergentes (127) trabalhou-se com a população total. Como o número de documentos onde houve concordância de julgamentos era maior (880), foi extraída uma amostra para o estudo, utilizando-se uma tabela de números aleatórios (201 documentos que usuário e não-usuário avaliaram *N* e 63 documentos que ambos avaliaram *M*).

Cada referência foi analisada no sentido de detectar as palavras do título que fossem iguais ou semanticamente relacionadas aos descritores do perfil. Tentou-se relacionar essa ocorrência comum de palavras com alguma propriedade que estivesse relacionada com a avaliação *M* ou *N*.

## 6 COMPARAÇÃO ENTRE JULGAMENTOS DE RELEVÂNCIA

A primeira hipótese aventada neste trabalho propõe uma comparação entre os julgamentos de relevância do usuário e do não-usuário para verificar a compatibilidade desses julgamentos.

### 6.1 Análise dos Dados e Resultados

O total de documentos avaliados (2033) foi distribuído em uma tabela (Tab. I) de acordo com o grau de avaliação que lhes foi conferido tanto pelo usuário como pelo não-usuário. Através de uma representação esquemática dessa tabela podemos compreender os dados anotados e a terminologia usada para se referir a eles.

As casas representadas em negrito evidenciam o número de documentos onde houve concordâncias de julgamentos.

Para se referir aos vários quadros da tabela usou-se sempre uma notação dupla (*MM*, *MR*, *MN* etc.) onde, a primeira letra sempre se refere ao julgamento do usuário e a segunda, ao do não-usuário.

		Julgamentos de Relevância do Não-Usuário		
		<b>M</b>	<b>R</b>	<b>N</b>
J u l g a m e n t o s	<b>d</b>	<b>MM</b>	<b>MR</b>	<b>MN</b>
	<b>e</b>	<b>RM</b>	<b>RR</b>	<b>RN</b>
	<b>u</b>	<b>NM</b>	<b>NR</b>	<b>NN</b>

Total

Tab. I

Tabela das avaliações

		Não-Usuário		
		M	R	N
U s u á r i o	M	210	263	110
	R	91	281	260
	N	17	131	670
<b>Total</b>				<b>2033</b>

Assim, como se pode observar, 210 *documentos foram* considerados *M* pelo usuário e também pelo não-usuário; 91 foram considerados *R* pelo usuário e *M* pelo não-usuário; e assim por diante.

Aconteceu, regularmente, que a divergência entre dois níveis (*M* e *R*, por exemplo) foi maior no sentido usuário—não-usuário que no inverso. O que parece indicar que a tendência do não-usuário foi sempre a de avaliar as referências abaixo do interesse que os documentos pudessem ter para a questão do ponto de vista do usuário. Uma avaliação de "interesse" e "não-interesse" simplesmente, talvez aproximasse mais os julgamentos.

Apenas 6, 25% dos documentos analisados mostraram *diferenças extremas de avaliação (MN e NM)*. E em 57, 10% dos casos houve concordância de julgamentos entre os usuários e o não-usuário.

Nem sempre os documentos considerados *M*, *R* ou *N* pelos usuários foram os mesmos considerados *M*, *R* ou *N*, respectivamente, pelo não-usuário. Assim, dos 28, 68% dos documentos considerados *M* pelos usuários, 12, 94% foram considerados *R* pelo não-usuário.

Foram analisados também os totais das avaliações *M*, *R* e *N* feitas pelo usuário e pelo não-usuário, sem implicar as diferenças individuais de avaliação para cada documento, isto é, sem que os documentos considerados *M* pelo não-usuário (3181, por exemplo, façam parte dos 583 documentos julgados *M* pelos usuários.

Esses dados foram reunidos numa tabela juntamente com os documentos em que houve concordância de julgamentos. (Tab. II).

Tab. II

Documentos em que houve concordância em relação ao total das avaliações

			Não-Usuário		
			M	R	N
			318	675	1040
U s u á r i o	M	583	210		
	R	632		281	
	N	818			670

Observe-se que dos 583 documentos julgados *M* pelos usuários e dos 318 julgados *M* pelo não-usuário, 210 foram julgados *M* por ambos e assim por diante.

## 6.2 Cálculo das Medidas Propostas

As medidas de sensibilidade ("sensitivity") e especificidade ("specificity") foram redefinidas para fins deste estudo para:

$$Se = \frac{\text{n}^{\circ} \text{ de documentos julgados } M \text{ por ambos}}{\text{n}^{\circ} \text{ de documentos julgados } M \text{ pelos usuários}}$$

Desta forma, o total de documentos julgados *M* ou *N* pelos usuários representa, respectivamente, o total de documentos relevantes ou irrelevantes do arquivo analisado.

Os resultados foram os seguintes:

$$Se = \frac{210}{583} = 0,36$$

$$Sp = \frac{670}{818} = 0,81$$

Normalmente, o nível de especificidade é bem maior que o de sensibilidade. Os resultados obtidos, vêm confirmar esse fato.

Em geral, é mais fácil para um substituto do usuário concordar sobre o que não é de interesse para a questão proposta pelo usuário do que sobre o que é de muito interesse. A medida do grau de interesse (muito, relativo) é subjetiva e varia de acordo com necessidades e pontos de vista do usuário no momento da avaliação.

## 6.3 Considerações sobre os Resultados

Através da análise dos julgamentos de relevância verificou-se que há diferença entre os julgamentos do usuário e do não-usuário. No entanto, a diferença não foi significativa a ponto de comprometer o papel de um não-usuário ao avaliar documentos para as questões propostas.

Houve maior concordância sobre o que não é relevante do que sobre o que é relevante. Assim, 28, 68% dos documentos foram julgados *M* pelos usuários e só 10, 33% foram documentos em que houve concordância de julgamentos. Mas, de 40, 23% dos documentos considerados *N* pelos usuários, 32, 95% eram documentos em que houve concordância de julgamentos.

Esse resultado foi também encontrado na literatura. SARACEVICI<sup>(21)</sup>, por exemplo, concluiu que, de acordo com as medidas de avaliação utilizadas em seu experimento (sensibilidade e especificidade), a concordância sobre o que é relevante foi consideravelmente menor que a concordância sobre o que é não-relevante.

As conclusões apresentadas por SARACEVICI<sup>(22)</sup>, em outro trabalho, mostram que quanto menos um não-usuário sabe sobre uma questão, maior a tendência de julgar documentos relevantes. Afirma o autor que os grupos de pessoas não conhecedoras do assunto (por exemplo, especialistas de sistemas de recuperação da informação e bibliotecários) tendem a fazer julgamentos de relevância relativamente altos. No entanto, o autor afirma que se espera que a avaliação de relevância de documentos seja similar para fornecedores de informação (isto é, substitutos do usuário) com algum conhecimento do assunto e para os usuários (isto é, pessoas com extenso conhecimento do assunto).

De fato, essa afirmação foi comprovada neste estudo. Embora a avaliação tenha sido feita por um bibliotecário (não-usuário), a tendência foi emitir julgamentos de relevância baixos. Isto porque, no presente caso, a bibliotecária, pelo fato de ter tido um treinamento sobre as disciplinas relacionadas à energia nuclear, se enquadra na segunda observação de SARACEVICI. Isto é, o fato de ter conhecimento do assunto fez com que os julgamentos de relevância do não-usuário fossem similares aos dos usuários.

Considera-se que as discordâncias de julgamentos muitas vezes ocorreram por falta de uma definição da escala de avaliação adotada. Em alguns experimentos!<sup>(21, 3, 6)</sup> os resultados indicaram que os julgamentos de relevância são melhor manipulados quando os julgamentos são instruídos adequadamente.

## 7 ANÁLISE SEMÂNTICA

A segunda hipótese proposta para este estudo está relacionada ao conteúdo dos títulos de documentos como elemento influenciador dos julgamentos de relevância. A referência bibliográfica do documento era o único elemento disponível na ocasião da avaliação e, assim, a avaliação foi feita com base nos títulos.

### 7.1 Análise

A determinação de propriedades textuais nos documentos analisados, que pudessem testar a hipótese acima, foi feita da seguinte forma:

- 1) Os 2033 documentos analisados foram agrupados de acordo com concordâncias e discordâncias de julgamentos do usuário e não-usuário.

- 2) Foram escolhidos para a análise os grupos de documentos onde houve concordância de julgamentos (MM e NN) e onde houve discordância extrema (MN e NM).
- 3) Os títulos dos documentos de cada grupo foram comparados com os descritores dos perfis para os quais haviam sido recuperados.
- 4) Fez-se uma tabulação dos dados de cada grupo de documentos de acordo com os números de descritores do perfil que apareceram nos títulos.
- Quatro grupos foram encontrados, variando de nenhum descritor a três descritores do perfil no título. (Tab. III).

Tab. III

Distribuição dos documentos de acordo com ocorrência de descritores

		Ocorrências de Descritores				Total
		a (nenhum descritor)	b (um descritor)	c (dois descritores)	d (três descritores)	
Grupos de Documentos	NM	3	13	1	0	17
	MN	58	45	7	0	110
	MM	10	29	21	3	63
	NN	132	67	2	0	201
						391

- 5) Comparou-se os resultados para cada grupo de documentos, definindo-se, então, as características de cada grupo e a ocorrência das propriedades a, b, c, d (nº de descritores do perfil no título) em relação às avaliações M, R e N.

7. 2 Resultados

Os Resultados da análise mostraram que o número de descritores do perfil que aparecem nos títulos dos documentos recuperados não é um fator decisivo para a avaliação de um documento como sendo relevante ou não-relevante.

Verificou-se, contudo, que os documentos considerados relevantes tinham mais descritores do perfil que os não-relevantes.

Observou-se que 65, 67% dos títulos dos documentos considerados como não-relevantes, tanto pelo usuário como pelo não-usuário, não continham nenhum descritor selecionado pelo perfil.

Por outro lado, 84, 12% dos documentos do grupo MM tinham um ou mais descritores do perfil no título, sendo que 38, 09% dos documentos tinham 2 ou mais descritores.

Pôde-se observar que, para o julgamento do não-usuário, o aparecimento de descritores do perfil nos títulos dos documentos influencia o julgamento e justifica e autoriza o não-usuário a classificar os documentos em relevantes e não-relevantes.

Quanto à relação entre os descritores contidos nos títulos, verificou-se que a maior parte dos

documentos que causaram divergências de julgamentos, e que continham ao menos um descritor de perfil, não mostrou o relacionamento entre os descritores indicado na questão.

Enquanto se fazia a análise dos resultados, notou-se que determinados perfis provocaram maiores discordâncias ou concordâncias de julgamentos que outros. Essa observação poderia conduzir a um ponto de vista mais crítico na interpretação dos resultados. Esse fato parece indicar que não só as características textuais dos documentos devem ser levadas em conta para a análise dos documentos, como também as flutuações de pontos de vista do usuário que ocorrem no momento da avaliação.

### 7.3 Considerações sobre os resultados

Os resultados alcançados indicaram que a frequência de descritores do perfil no título não discrimina, por si só, os documentos relevantes dos não-relevantes.

Verificou-se, realmente, que os documentos relevantes tinham mais descritores do perfil no título que os não-relevantes. Mas essa verificação não deu condições para que se estabelecesse um padrão de frequência de descritores nos documentos relevantes ou nos não-relevantes, que designasse um padrão relativo de julgamentos de relevância.

Embora as condições e elementos de estudo não tenham sido os mesmos dos experimentos realizados por TAGUE<sup>(22)</sup>, GIFFORD e BAUMANIS<sup>(6)</sup> e IVANKIN<sup>(8)</sup>, as condições foram compatíveis.

## 8 CONCLUSÕES

De acordo com os resultados obtidos para o teste da primeira hipótese proposta, concluiu-se que há discordâncias entre os julgamentos do usuário e do não-usuário. Essa diferença, contudo, não foi significativa a ponto de comprometer o papel de um não-usuário ao avaliar documentos, em substituição ao usuário.

Considera-se que o fato de não haver ocorrido um índice de concordância mais alto se deve ao fato de que, enquanto o usuário avalia os documentos para sua necessidade de informação, o não-usuário os avalia para a questão proposta. Essa conclusão poderia ser testada se se instrísse os usuários no sentido de formular a questão o mais próximo possível de sua necessidade de informação e de avaliar os documentos sob o ponto de vista exclusivo da questão proposta.

A segunda hipótese relaciona a frequência de aparecimento de descritores do perfil nos títulos de documentos aos julgamentos de relevância. Conclui-se que o número de descritores do perfil nos títulos não é um fator determinante de relevância de documentos.

Em relação aos serviços de SDI e ao entendimento da conexão entre títulos de documentos e julgamentos de relevância podemos concluir que:

- 1) O não-usuário com algum conhecimento do assunto e familiaridade com o enfoque das questões pode avaliar documentos ou questões em substituição ao usuário.

Deve-se reservar espaço adequado no formulário de proposição da questão para que o usuário esclareça o assunto envolvido na questão, a finalidade, as circunstâncias determinantes do estudo e o tipo de informação desejada.

- 2) Deve-se criar condições que façam o usuário interagir com o sistema.
- 3) Deve-se definir para o usuário a escala de avaliação adotada.
- 4) O aparecimento ou não de descritores do perfil nos títulos dos documentos é para o não-usuário um critério de julgamento.
- 5) Deve-se procurar, tanto quanto possível, relações semânticas entre descritores, desde que essas relações não signifiquem um conceito já adotado no thesaurus.

Uma idéia sugerida, mas não testada, durante este estudo foi que quanto mais detalhada a escala de avaliação maior o índice de discordâncias de julgamentos entre usuário e não-usuário. Pareceu relativamente fácil para o não-usuário concordar que um documento seja relevante ou não-relevante face à questão. O mesmo não aconteceu, entretanto, quanto à determinação do grau de relevância (muito, relativo, etc.) do documento em relação à necessidade de informação do usuário.

Além disso, do ponto de vista do sistema existe uma avaliação dicotômica: relevante e não-relevante; do ponto de vista do usuário existem diferentes graus de relevância dos documentos.

## CITAÇÕES BIBLIOGRÁFICAS

- 1 - BARHYDT, G. C. The effectiveness of non-user relevance assessments. *J. Docum.*, 23(2): 146-9, e (3): 251, 1967.
- 2 - BRADFORD, S. C. Sources of information on specific subjects. *Engineering*: 85-6, 1934 apud FIGUEIREDO, L. M. de. O conceito de relevância e suas implicações. *CL Inf.*, 6(2): 75-8, 1977.
- 3 - CUADRA, C. A. & KATTER, R. V. Opening the black box of "relevance". *J. Docum.*, 25(4): 291-303, Dec. 1967.

- 4 - FOSKETT, D. J. A note on the concept of "relevance". *Inf. Storage Retrieval*, 8(2): 77-8, Apr. 1972.
- 5 - GEBHARDT, F. A simple probabilistic model for the relevance assessment of documents. *Inf. Process. Mgmt*, 77(1/2): 59-65, Jun. 1975.
- 6 - GIFFORD, C. & BAUMANIS, G. J. On understanding user choices: textual correlates of relevance judgements. *Am. Docum.*, 20(1): 21-6, Jan. 1969.
- 7 — HARMON, G. Information need transformation during inquiry: a reinterpretation of user relevance. *Proc. ASIS*, 7: 41-3, 1970.
- 8 - IVANKIN, V. I. A study of the semantic relationships between the texts of a query and of a relevant document. *Autom. Docum. math. Linguistics*, 9(1): 20-7, 1975.
- 9 — KENT, A. et alii. Machine Literature searching. VIII. Operational criteria for designing information retrieval systems. *American Documentation*, 6(2): 93-101, 1955 apud SARACEVIC, T. Relevance: a review and framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.*, 26(6): 321-43, Nov. /Dec. 1975.
- 10 — KING, D. W. Design and evaluation of information systems. *A. Rev. Inf. Sci. Technol.*, 3: 70-4, 1968.
- 11 - LANCASTER, F. W. Some notes on the distinction between pertinence and relevance. In: \_\_\_\_\_ Guidelines for the evaluation of information systems and services, (s. I.), (s. ed.), Jan. 1977. (Prepared for Unesco under contract).
- 12 - MOOERS, C. S. Coding, information retrieval, and the rapid selector. *American Documentation*, 7(4): 225-9 (1950) apud SARACEVIC, T. Relevance: a review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.*, 26(6): 327, Nov. /Dec. 1975.
- 13 - NATIONAL ACADEMY OF SCIENCES. Proceedings of the International Conference on Science Information, 2 Vols. National Academy of Sciences, Washington D. C. (1959) apud SARACEVIC, T. Relevance: a review of and a framework for the thinking of the notion in information science.: *J. Am. Soc. Inf. Sci.*, 26(6): 327, Nov. /Dec. 1975.
- 14 — O'CONNOR, J. Relevance disagreements and unclear request forms. *Am. Docum.*, 18(3): 165-77, 1967.
- 15 — \_\_\_\_\_. Some independent agreements and resolved disagreements about answer-providing documents. *Am. Docum.*, 20(4): 311 -9, 1969.
- 16 - PERRY, J. W. Superimposed punching of numerical codes on handsorted, punch cards. *American Documentation*, 2(4): 205-212 (1951) apud SARACEVIC, T. Relevance: a review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.*, 26(6): 327, Nov. /Dec. 1975.
- 17 - POLUSHKIN, V. A. Relevance and pertinence. *Autom. Docum. math. Linguistics*, 7(1): 52-4, 1973.
- 18 — RATH, G. J. et alii. Comparisons of four types of lexical indicators of content. *Am. Docum.*, 12(2): 126-30, 1961.
- 19 - RESNICK, A. & SAVAGE, T. R. The consistency of human judgements of relevance. *Am. Docum.*, 75(21): 93-5, 1964.
- 20 - SARACEVIC, T. Relevance: a review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.*, 22(2): 126-39, Mar. /Apr. 1971.
- 21 — \_\_\_\_\_. Selected results from an inquiry into testing of IR systems. *J. Am. Soc. Inf. Sci.*, 22(2): 126-39, Mar. /Apr. 1971.
- 22 — TAGUE, J. Matching of question and answer terminology in an educational research file. *Am. Docum.*, 76(1): 26-32, Jan. 1965^
- 23 - TAUBE, M. et alii. Storage and retrieval of information by means of the association of ideas. *American Documentation*, 6(1): 1-17 (1955) apud SARACEVIC, T. Relevance: a review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.*, 26(6): 327, Nov. /Dec. 1975.
- 24 - WILSON, P. Situational relevance. *Inf. Storage Retrieval*, 9(8): 457-71, 1973.

#### ABSTRACT

User and non-user relevance judgments of a group of documents in the field of nuclear energy, were compared. Assessments of relevance were done by documents titles. The effectiveness of titles in determining relevance or non relevance of documents was observed. It was concluded that user and non-user judgements were similar with an acceptable number of disagreements. It was considered that the level of agreements was not higher because users evaluate documents to their information needs, while non-users evaluate them to the proposed question. The frequency of profile descriptors in document titles does not differentiate relevant documents from non-relevant ones.