REPRESENTAÇÃO INDEXAL NA *WEB:* estudo do sintagma "História da Paraíba" nos sites *Alta Vista* e *Google*[*]

THE INDEXAL REPRESENTATION IN THE WEB: a study of the syntagm "História da Paraíba " in the Alta Vista and Google sites

Joliza Chagas Fernandes[**]
Virginia Bentes Pinto[***]
Carlos Xavier de Azevedo Netto[****]

Resumo

Apresenta o resultado do estudo exploratório sobre a representação do sintagma nominal "História da Paraíba" na *WEB*, nos *sites Google* e *AltaVista*, objetivando investigar a sua forma de organização representacional, no contexto da indexação, e a cobertura de conteúdos, tomando como parâmetro

[*] Baseado na Dissertação de Mestrado de Joliza Chagas Fernandes

[**] Mestre em Ciência da Informação - Universidade Federal da Paraíba. Professora Assistente do Curso de Biblioteconomia da UFMT. Vice-líder do grupo de pesquisa "Estudos Avançados em Informação" da Universidade Federal de Mato Grosso - UFMT. jolizah@ufmt.br

[***] Doutora em Ciências da Informação e da Comunicação – Université Stendhal 3 – França. Profª da Universidade Federal do Ceará. vbentes@ufc.br; bentespinto@yahoo.com.br

[****] Doutor em Ciência da Informação – Universidade Federal do Rio de Janeiro. Professor do Departamento de Biblioteconomia e Documentação da Universidade Federal da Paraíba. xaviernetto@ig.com.br

os títulos da base de dados da USP, na área de "História da Paraíba". A metodologia adotada foi a busca direta no *Google* e no *AltaVista*, observação intensiva individual e entrevista não estruturada, além das técnicas da análise de conteúdo. Os resultados mostram que a representação indexal na *WEB* necessita de reajustes no que diz respeito à função de representação dos documentos eletrônicos para atingir as expectativas deste novo canal de informação. No que diz respeito à cobertura de conteúdo comprovou-se que, apesar de ter alcançado um percentual pequeno de completeza em ambos os *sites*, tomando por referência a base de dados DEDALUS da USP, os mesmos possuem assuntos de importância na área de História da Paraíba.

Palavras chaves

REPRESENTAÇÃO DA INFORMAÇÃO
INDEXAÇÃO
RECUPERAÇÃO AUTOMÁTICA DA INFORMAÇÃO

1 INTRODUÇÃO

Indubitavelmente, a Internet é a maior rede mundial de informação e de comunicação da sociedade contemporânea, em cujo espaço estão embutidas possibilidades inimagináveis que vão desde a produção até o acesso e uso de informação. No entanto, mesmo abrigando uma grande variedade de provedores, programas e produtores de serviços, o acesso aos conteúdos de seus repertórios documentais, independentemente se buscadores ou diretórios, torna-se muitas vezes impraticável, frente ao emaranhado em que se encontra a informação disponível, tornando-a fluida e rarefeita diante do seu excesso de possibilidades "ofertadas" aos olhos humanos.

Nesse emaranhado de possibilidades se destaca a *WORD WIDE WEB*, popularmente denominada *WEB*. Como o principal repositório de informação da Internet, além de disponibilizar o seu grande conteúdo independentemente de tempo e espaço, também se torna alvo de especulação e preocupação de várias áreas de conhecimento, notadamente da Ciência da Informação, devido a sua ineficácia no processo de recuperação da informação. Pois, muitas vezes, transforma a busca numa turbulenta e impraticável recuperação informacional à medida que, quase sempre fornece uma avalanche de respostas não relevantes aos seus usuários. Vários estudos já foram desenvolvidos na perspectiva de investigar, entre outros, a organização e a recuperação da informação nesta rede. As conclusões desses estudos mostram que, a representação indexal das informações na Internet é organizada através de palavras-chave do tipo "unitermo" - neologismo bibliotecário criado por Mortimer Taube em 1953 para designar elementos unitários, ou palavras simples. Estes unitermos têm contribuído para aumentar o chamado ruído (*recall*) na recuperação.

A constatação destes fatos, no nosso cotidiano de trabalho é que nos motivou a empreender uma pesquisa junto ao Curso de Mestrado em Ciência da Informação da Universidade Federal da Paraíba, com objetivo básico de analisar a forma de organização representacional na *WEB*, especificamente dos repertórios documentais brasileiros abrigados nos *sites AltaVista* e *Google*, considerados como os mais consultados pela comunidade virtual brasileira, segundo mostram as estatísticas fornecidas pela *PageRank*. Para alcançar este objetivo, partimos dos seguintes questionamentos: Como estão organizadas e representadas as informações no contexto da indexação dos *sites Google* e *AltaVista*? Durante o processo de busca as respostas ofertadas são pertinentes e eficazes?

2 CONSIDERAÇÕES SOBRE A REPRESENTAÇÃO

2.1 REPRESENTAÇÃO INDEXAL

O termo representação, em sentido amplo, remonta aos pré-socráticos, principalmente no campo da filosofia, da matemática e da

religião. A partir do século XVII, volta a ser destaque, entre outros com Kant e, no século XX com Peirce, Frege, Wittgenstein e Saussure. Mais recentemente, a partir dos anos 1940 ele passa a ser estudado, notadamente no domínio das ciências cognitivas: Psicologia, Lingüística, Sociologia e Ciência da Informação, onde os estudos ainda são escassos. Nestes contextos, fala-se, entre outros conceitos, de representação social, representação do conhecimento e representação da informação. Independentemente do campo de conhecimento, a compreensão da representação traz implícita a noção de um elemento estar em lugar de outro, como bem trabalha Peirce (1977, p.61), ao expressar que "representar é estar em lugar, isto é, estar numa tal relação com o outro que, para certos propósitos, é considerado por alguma mente como se fosse outro". No *Grand dictionnaire de la psychologie* (1991, p.596) este conceito pode designar, o processo que coloca em correspondência pelo menos dois elementos, "de forma que um (o representante) repita, substitua ou apresente de outra forma o outro (o representado). Também designa um aspecto do resultado desse processo em ocorrência o único elemento representante, de qualquer natureza que ele seja". Entendemos que estas propostas vêm ao encontro, entre outras, das áreas da Ciência da Informação e da Biblioteconomia, pois, desde a sua estruturação se apropriam da representação sobre vários aspectos, destacando-se o processamento, a recuperação, a comunicação, a recepção e uso da informação. Conforme Bentes Pinto; Mota; Queiroz (2003), ao se processar informações, catalogando, indexando ou estabelecendo estratégias de busca, "necessariamente, trabalha-se com signos, sejam eles verbais ou não verbais, com um fim determinado: contribuir para o acesso a informações, mesmo que em muitos casos isto nem sempre seja possível". Entretanto, Azevedo Netto (2002 *apud* PEIRCE, 1977, p.3), diz que na compreensão sígnica está implícito o reconhecimento da possibilidade de construção de significados, uma vez que

> [...] esta entidade, antes de estar no lugar de alguma coisa, ou mesmo representa-la é um processo de encadeamento, no qual o signo leva a construção de um outro signo que se relaciona com o primeiro, e assim por diante em uma constante

> semiose. Este fato deve-se à identificação de uma das entidades que compõe o signo, o interpretante, que possui em si o atributo de produção de significado [...] O signo é um feixe de relações em que ocorre uma relação triádica entre o objeto, veículo e interpretante, dentro da construção da significação e do processo de comunicação.

Ora, como dito anteriormente, na representação indexal, também se mexe com signos, principalmente palavras, visando a construção de índices, tanto no momento do tratamento da informação pelos indexadores, como pelos usuários, nas definições das estratégias de busca da informação. Na construção desses índices, principalmente automáticos, adota-se quase sempre unitermos, que na visão de Bentes Pinto (2001, p.226), acarretam um empobrecimento do poder de representação dos conteúdos dos documentos indexados. Esta maneira de indexar

> retira as palavras do contexto do lógico semântico onde elas tinham uma significação determinada por este contexto. Elevadas do mundo real, tais palavras designam um conjunto de propriedades, e seu sentido muda resumindo-se a um conjunto de unidades léxicas.

Corroborando Le Guern (1991, p.23) diz que, as palavras da língua, ao contrário de suas ocorrências no discurso, são desprovidas de

> referência extralingüística. [...]. A relação signo/objeto, ao senso de Peirce, corresponde a relação significante significado; os significantes fazem parte também da estrutura da língua, significam somente propriedades, nunca entidades; elas significam atributos, e não substância [...].

Ainda neste escopo, Azevedo Netto (2002, p.57), mostra que

> o processo de significação se dá através do interpretante [...] Assim, o significado pode ser vis-

> to como uma construção, que varia nos contextos de interlocução e dentro de instâncias culturais distintas, propiciando a elaboração de interpretações sobre o que está sendo representado.

2.1.1 Representação indexal na Net

No ambiente virtual, entendemos que a representação informacional tem suas origens nas primeiras experiências de indexação, com uso de máquinas, por volta da década de 1959, quando H.P. Luhn colocou em prática o índice KWIC (*key-world in context*). Na década de 1970, apareceram outras propostas, destacando-se o SMART desenvolvido por G.Salton (1971) e, tendo por base os modelos estatísticos e probabilísticos de ocorrência e co-ocorrência de palavras. Estas experiências de uso das tecnologias no tratamento e busca da informação, a exemplo da indexação manual, também foram estruturadas tendo como base os unitermos e palavras-chave, tanto para a representação indexal como nas estratégias de busca, o que contribui para revocação das respostas às demandas dos usuários.

Com o aparecimento da Internet e a estruturação da *WEB*, constituída por uma infinidade de documentos eletrônicos com conteúdos sobre os mais diversos assuntos, observamos, mais uma vez que, a representação indexal e a recuperação da informação são automáticas, e estruturadas fundamentalmente por unitermos, atribuídos pelas máquinas no momento das buscas. Portanto, de "forma linear e determinista, e, certamente, não atenderão eficazmente às necessidades informacionais de seus usuários, uma vez que cada indivíduo possui uma maneira de representar e organizar a informação, o que nem sempre coincide com a forma da *WEB"* (FERNANDES, 2004). Essa organização também é baseada em modelos estatísticos e probabilísticos de ocorrência e co-ocorrência de unitermos, que muitas vezes não representam os temas abrigados nos sites, fornecendo respostas com conteúdos exaustivos e sem relevância para o usuário. Isto acontece porque as ferramentas de busca *on-line* oferecem como respostas, além dos termos demandados, outros com raízes idênticas, porém com suas variações gramaticais e

significados diferentes, provocando, mais uma vez, ruídos informacionais, demonstrando que esse espaço é, por natureza, entrópico. Em outras palavras, o que se percebe é que, na realidade, se criou um caos estruturado na *WEB*, cujos objetos, signos e significados se tornam elementos confusos para o usuário no momento da recuperação da informação. Diante disto, constata-se mais do que nunca, que as técnicas de tratamento, organização e recuperação da informação são instrumentos de fundamental importância que podem contribuir para a reestruturação desse espaço, através do tratamento criterioso dos itens de informação, visando a sua recuperação *a posteriori*.

Neste sentido é que está sendo ventilada, desde 2000, uma nova proposta para a *WEB*, denominada *WEB* Semântica. Nesta nova versão da *WEB*, os conteúdos da rede serão organizados e representados, levando-se em consideração a semântica dos dados, isto com a intenção de facilitar a busca e a recuperação de informações em níveis contextuais. Portanto, já se percebe a necessidade de um processamento informacional das fontes de informação da rede no que diz respeito a sua representação nos processos de indexação, pois, como chama atenção Chaumier (1990), de nada serve arquivar um documento se não se sabe encontrá-lo porque ele não foi indexado ou, ainda, se ele foi indexado de maneira incorreta. Na mesma linha de preocupação, Lawrence; Giles (1999, p.15), mostram o desperdício de tempo ao se buscar informações na rede Internet. Os "usuários da web gastam muito tempo usando ferramentas de busca ('search engines') para localizar material na vasta e desorganizada web [...] várias destas ferramentas estão, inequivocamente, classificados entre os 10 sites mais acessados da web". Corroborando Bougnoux (1993, p.11) argumenta que, nunca se teve tanta ilusão de estar tão bem informado, o que não quer dizer que se sabe tratar e integrar os dados que literalmente submergem, pois "muita informação pode matar a informação, suscita evasões imaginárias, e recusa de saberes, e se choca de qualquer maneira ao 'segredo informacional' de cada um (um organismo só utiliza uma ínfima parte dos sinais que perpassam pelo seu meio ambiente)".

Conforme argumenta Gimenez Lugo (2002), estas preocupações fazem parte do campo da Biblioteconomia e da Ciência da Informação, que sempre estiveram na vanguarda do processamento e orga-

nização da informação, na perspectiva, entre outras, de negar a entropia decorrente da explosão documental, e fornecer subsídios que possam auxiliar seus usuários nas buscas e recuperação de itens de informação. Corroborando, Milstead (1999) afirma que, em relação ao tratamento e representação dos recursos informacionais, os "bibliotecários e indexadores têm produzido e padronizado metadados por séculos". Ora, a representação em atividades de informação, tais como catalogação, classificação e indexação de documentos em suportes tradicionais, é tão comum, entre outras desenvolvidas pelos profissionais de informação, fora do âmbito da *WEB*, que, às vezes, pode até ser um tanto quanto despercebida. Porém, sempre teve função primordial para o processo de recuperação de informação.

3 METODOLOGIA

A pesquisa classifica-se como estudo exploratório, com intenção de avançar o conhecimento sobre a representação e a indexação na *WEB*, e também compreender as mudanças de paradigmas, em conseqüência do aparecimento dos aparatos tecnológicos de produção, tratamento, organização, disseminação e recepção da informação. A coleta de dados foi realizada nos *sites Altavista e Google*, e também na Base de dados DEDALUS, com observação direta individual e anotações no diário de campo. Para a análise de conteúdo, tomamos por base a proposta de Bardin (1977) e as categorias de análise de Lancaster (1993), quais sejam: a) cobertura: trata-se da abrangência de conteúdo na área de "História da Paraíba", dos *sites* estudados, b) previsibilidade na recuperação da informação: trata-se da possibilidade de se conhecer um item relevante a partir de algumas pistas sobre esses itens, contidas nos títulos das "chamadas dos *sites*", os títulos de itens foram denominados "títulos de documentos eletrônicos", c) recuperabilidade: trata-se da capacidade de um sistema de recuperação de informação recuperar itens relevantes em uma determinada busca. Afora estas, criou-se ainda as categorias de representação da superestrutura e ordenação lógica. A primeira diz respeito à análise "representacional" dos documentos

indexados nos *sites* pesquisados, levando-se em consideração a linguagem de indexação utilizada – linguagem natural ou documentária. A ordenação lógica trata da hierarquia de ordenação dos resultados em uma busca, ou seja, o porquê de uma página de determinado *site* aparecer como o primeiro item da listagem recuperada ou o último. Também, utilizamos a entrevista não estruturada com sete professores e pesquisadores da UFPB e da UFMT, da área de história. Esta entrevista foi realizada com objetivo de se conhecer qual fonte de informação secundária era mais utilizada pelos referidos professores, para que pudéssemos comparar a eficácia entre a recuperação da informação manual e automática.

Durante a fase de pré-teste nos *sites* e na base DEDALUS utilizamos várias estratégias de busca: História da PARAÍBA, "História da Paraíba", Paraíba - História, Paraíba, História. Estas estratégias nos forneceram um "calhamaço" de documentos que na maioria não diziam respeito ao assunto "História da Paraíba", justamente porque a indexação no espaço *ciber* tem por base o "unitermo". Então, para a pesquisa definitiva, optamos pelo uso do sintagma "História da Paraíba" em uma busca fechada, ou seja, entre aspas.

4 ANÁLISE E INTERPRETAÇÃO DOS DADOS

Após a coleta dos dados, passamos a analisá-los tendo por base as seguintes categorias: recuperabilidade, previsibilidade, representação da superestrutura, ordenação lógica dos resultados e a cobertura dos *sites* estudados.

4.1 RECUPERABILIDADE NO GOOGLE

O *Google*, em primeira instância, provou, em números, que realmente é um dos mais expressivos buscadores, resgatamos 3.522 itens sobre o sintagma "História da Paraíba". No entanto, observando-se os resultados das buscas diárias, chegamos apenas a 148 itens para pesquisa. De posse destes itens, fizemos a sua inspeção. Durante esta fase,

percebemos que havia páginas que não eram encontradas, ou por não se encontrarem mais na rede, ou por alguma falha ou erro de programação (erro 404). Mediante estes resultados, eliminamos os itens não encontrados, e selecionamos os que conseguíamos acessar e que se revelavam importantes para o estudo. Dentro desta seleção, foram observados os itens relevantes e os irrelevantes ao tema proposto, conforme Gráfico 1.

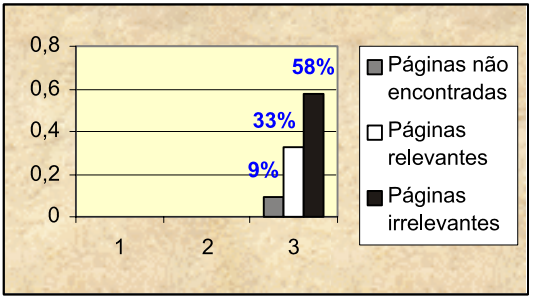

Gráfico 1: Expressões Informacionais *Google*
Fonte: Pesquisa direta **www.goolge.com.br** em outubro de 2003

Com esta avaliação, chegamos a 03 (três) resultados: 13 registros concernentes a páginas não encontradas (erro 404), ou seja, 9% do total de itens selecionados para a pesquisa; 86 registros irrelevantes, o que equivale a 58% dos itens; e apenas um total de 49 registros de itens ou páginas relevantes, perfazendo 33%. De posse desses, escolhemos para as análises somente os itens, ou páginas, considerados como relevantes e irrelevantes para a avaliação da recuperabilidade, ou seja, analisamos um total de 135 registros. Após a primeira filtragem, passamos a uma segunda, desta vez já com os elementos (itens ou páginas) selecionados. Com estes resultados, realizamos uma inspeção direta em cada item verificando o conteúdo de cada um, observando se continha ou não elementos sobre a História da Paraíba propriamente, ou seja, textos que incluíam artigos, apontamentos, notas, datas etc. Nas respostas obtidas nesta análise, percebemos a repetição de ocorrência da primeira análise, ou seja, as mesmas 49 páginas ou itens mantiveram-se relevantes. Estes resultados confirmam a falta de precisão, no referido *site*, a res-

peito da recuperação da informação, o que pode provocar um estresse na busca pelas informações desejadas, em detrimento da dificuldade de seleção, uma vez que há excesso de informações irrelevantes em meio a um reduzido percentual de informação relevante, provocando o que Lardy (1997) denomina alto índice de ruídos informacionais. Neste sentido, observamos que a navegação através do referido mecanismo de busca por palavra-chave, mesmo fechada (entre aspas), embora pareça simples, se revela conturbada, necessitando de uma filtragem mais minuciosa para obter um resultado eficaz.

4.2 RECUPERABILIDADE NO ALTAVISTA

Na pesquisa do *AltaVista*, resgatamos 297 itens sobre o sintagma "História da Paraíba" durante o tempo de coleta, mas quando das buscas diárias, obtivemos como resultado apenas 41 itens. De posse destes itens, passamos à inspeção, constatando que havia páginas que não eram encontradas, ou por não se encontrarem mais na rede, ou por alguma falha ou erro de programação (erro 404). Então, eliminamos os itens não encontrados, [illegible] er acessados, ou seja, 41 itens relevantes [illegible] Com esta análise, chegamos a 03 (três) [illegible] encontradas, na percentagem de 13%. [illegible] relevantes foram encontradas 16, o que [illegible] a categoria. Já as respostas das páginas [illegible] gistros, ou seja, (45%). Vede o Gráfico 2.

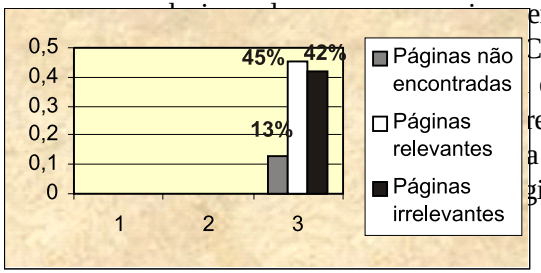

Gráfico 2: Expressões Informacionais *AltaVista*
Fonte: Pesquisa direta **www.altavista.com.br** em outubro de 2003

Com os dados da primeira filtragem, tomamos para as análises somente os itens, ou páginas, considerados como relevantes e irrelevantes para a avaliação da recuperabilidade, ou seja, analisamos um total de 33 registros. Após a primeira filtragem passamos a uma segunda, sendo feita a inspeção direta em cada item. Verificamos o conteúdo de cada um, observando se continha ou não elementos sobre a História da Paraíba propriamente, ou seja, textos que incluíam artigos, apontamentos, notas, datas etc. Nesta análise, percebemos a repetição de ocorrência da primeira análise, portanto, os mesmos 17 itens anteriormente mantiveram-se relevantes. Os termos relevantes ou pertinentes ao assunto perfazem um percentual considerável em relação às respostas da primeira busca, quase 50 % de relevância. Esta realidade pode ser em razão de que o *AltaVista* é, além de índice, como o *Google*, também um diretório, que usufrui do beneficio de contar com a presença de humanos para efetivar a organização da informação antes de abrigar as páginas na sua base de dados.

4.3 COTEJAMENTO DA RECUPERABILIDADE GOOGLE E ALTAVISTA

Em que concerne a este item, ficou evidente que, em ambos os *sites*, houve a deficiência de precisão na recuperação. No entanto, apesar de apresentar uma precisão baixa, o *AltaVista* se mostrou mais preciso, e, embora tenha recuperado menos itens em relação ao *Google*, foi mais objetivo em seus resultados, conforme Tabela 1.

Tabela 1: Recuperabilidade *Google* e *AltaVista*

Fonte: Pesquisa direta em outubro de 2003

As respostas mostram que o *AltaVista* conseguiu ser mais contundente em seus resultados, economizando um pouco mais de tempo do usuário que navega e precisa da informação num menor espaço de tempo. Como referido anteriormente, isto pode decorrer da presença de humanos na gestão informacional do *site.*

4.4 PREVISIBILIDADE NA RECUPERAÇÃO DA INFORMAÇÃO

Nesta categoria, tomamos por base as chamadas de cada item recuperado nas telas de respostas, os títulos dos documentos eletrônicos, os resumos e o texto integral dos referidos documentos nos *sites Google* e *AltaVista.* Desta forma, verificamos as pistas que cada um trazia nos elementos constituintes de sua macro e microestrutura, buscando todos os pontos de acesso que pudessem levar ao texto integral. Fazendo análise somente das chamadas de cada item apresentado na tela de resultados, verificamos que, dentre as 135 páginas ou itens recuperados, apenas 19 continham informações que auxiliavam na escolha de itens possivelmente relevantes, enquanto o restante não expressou indícios sobre o conteúdo, levando-nos a concluir que poderiam ser itens irrelevantes para a pesquisa. De posse dos resultados da análise das chamadas, efetuamos uma segunda análise para ver a eficácia junto aos títulos dos documentos a que elas se referem. Dos mesmos 135 itens recuperados, 22 portavam algumas pistas contendo o sintagma "História da Paraíba", os quais consideramos como indicativo de relevância do documento. Com relação a esta análise, constatamos que as pistas apresentadas nas chamadas, juntamente com as dos títulos, constituíam pistas mais eficazes de acesso ao documento. Neste sentido, Lancaster (1993) diz que os títulos de periódicos, como documento secundário, entendidos aqui como chamadas de *homepage*, isoladamente, não refletem muito o seu conteúdo, sendo necessário observar, juntamente com estes, o título do artigo, que é a obra primária, aqui representada pelo documento eletrônico. No que diz respeito ao resumo analisamos aqueles apresentados nas chamadas de cada *homepage,* fazendo tam-

bém uma relação destes com os títulos dos documentos eletrônicos integrais. Nesta análise, observamos que os resumos utilizados são constituídos de extratos do próprio texto integral, numa linguagem natural. O resultado foi menor do que aqueles obtidos na análise do título, ou seja, dos 135 itens recuperados, somente 21 portavam algumas pistas que representavam o conteúdo do documento.

O último estudo concernente à categoria previsibilidade foi realizado no texto integral, como previsto anteriormente, ou seja, verificando todos os 135 documentos recuperados. O resultado deste estudo apresentou maior número de itens relevantes da busca em relação às outras análises desta categoria, perfazendo um total de 49 registros. Nesta situação, as previsões de relevância melhoraram consideravelmente, comparando-se aos resultados obtidos nas análises das chamadas das *homepages*, dos títulos de documentos e dos resumos, o que confirma as reflexões de Lancaster (1993), ao anotar que, relativamente à extensão da representação do documento, quanto maior for tal representação, maiores serão as possibilidades de previsões para recuperação da informação, mesmo que o ruído seja considerado elevado, pois o usuário terá muito mais possibilidades para a seleção de itens relevantes. Para melhor visualização, mostramos os dados quantitativos destas análises na Tabela 2.

Tabela 2: Previsibilidade *Google*

CATEGORIAS	Nº DE ITENS RECUPERADOS	Nº DE ITENS CONSIDERADOS POSSIVELEMENTEIRR ELEVANTES	Nº DE ITENS CONSIDERADOS POSSIVELMENTE RELEVANTES
Chamadas da Homepage	135	116	19
Titulo do documento eletrônico	135	113	22
Titulo mais resumo da homepage	135	114	21
Texto integral do documento	135	86	49

Fonte: Pesquisa Direta em outubro de 2003

Com relação à previsibilidade do *site AltaVista*, adotamos também a inspeção direta de cada item recuperado, a fim de verificar as pistas que eles traziam nos elementos constituintes de sua macro e

microestrutura, ou seja, todos os pontos de acesso que pudessem levar ao texto integral. Inicialmente, analisamos somente as chamadas de cada item apresentado na tela de resultados. Nela, percebemos a existência ou não de elementos significativos, ou pistas que levassem ao documento contendo o conteúdo relevante sobre História da Paraíba. Para tanto, examinamos as 33 *homepages* recuperadas no *AltaVista* quando da primeira filtragem para a análise das categorias. Apenas 4 apresentaram títulos contendo informações que auxiliem na escolha de itens possivelmente relevantes, pois o restante consideramos como itens irrelevantes para a pesquisa. Estes resultados refletiram aqueles obtidos no *Google* e vão ao encontro das reflexões de Volpato (2002) quando argumenta a necessidade de se utilizar títulos menos polissêmicos, pois estes podem ser pistas mais eficazes para se recuperar os documentos a que dão título.

De posse destes resultados, realizamos a segunda análise, desta vez para ver a eficácia junto aos títulos dos documentos a que elas se referem. Dos 33 itens recuperados, apenas 13 portavam algumas pistas contendo o assunto História da Paraíba, as quais consideramos como indicativos de relevância do documento. Estes resultados mostram que as pistas apresentadas nas chamadas, juntamente com as dos títulos, constituíam pistas mais eficazes de acesso ao documento, o que ratifica o pensamento de Lancaster (1993), ao destacar que o título isoladamente pode não remeter ao conteúdo de um documento no momento da busca e recuperação da informação.

No que concerne ao resumo, analisamos aqueles apresentados nas chamadas de cada *homepage*, fazendo também uma relação destes com os títulos dos documentos eletrônicos integrais. Neste *site*, também observamos que os resumos são constituídos de extratos do texto integral, em linguagem natural. Constatamos um resultado melhor do que aqueles obtidos na análise do título, ou seja, dos 33 itens recuperados, 15 portavam algumas pistas que representavam o conteúdo do documento, aproximadamente 46% de relevância, percentual considerável em se tratando do ambiente hipermídia. Tal fato pode ser decorrente da presença de humanos na gestão informacional deste *site*, o que também foi percebido na análise da categoria recuperabilidade. Isto evidencia a importância do humano nas atividades de produção e organização

da informação, e nos remete à impossibilidade de substituição da subjetividade humana pela máquina.

A análise do texto integral foi o último elemento estudado na categoria previsibilidade do *site AltaVista*, como previsto, ou seja, verificando todos os 33 documentos recuperados. Como no *Google*, também apresentou como resultado um maior número de itens relevantes da busca em relação às outras análises desta categoria, perfazendo um total de 17 registros.

Como se pode observar, as previsões de relevância também foram bastante positivas se comparadas àqueles resultados obtidos nas análises das chamadas das *homepages*, dos títulos de documentos e dos resumos. Estes resultados podem ser visualizados melhor na Tabela 3.

Tabela 3: Previsibilidade *AltaVista*

CATEGORIAS	Nº de itens apresentados	Nº de itens considerados possivelmente irrelevantes	Nº de itens considerados possivelmente relevantes
Título da homepage	33	29	04
Título do Documento eletrônico	33	20	13
Título mais resumo da homepage	33	18	15
Texto integral do documento	33	16	17

Fonte: Pesquisa Direta em outubro de 2003

No cotejamento da previsibilidade *Google* e *AltaVista*, percebemos que, de maneira geral, em ambos os *sites*, quanto maior a extensão da representação das informações, maior a previsão de se encontrar documentos ou itens relevantes. Nos *sites* pesquisados, foi necessária uma consulta ao texto integral para se conseguir identificar os itens realmente relevantes, uma vez que as chamadas da *homepage* e o título do documento não contribuíram de maneira eficaz para esta previsão, pois não trouxeram informações que servissem de indicações sobre o conteúdo do documento.

Com relação à análise dos resumos, percebeu-se uma diferença no comportamento dos *sites*, ou seja, embora o *AltaVista* não tenha oferecido elementos suficientes para a previsão de todos os itens relevantes da busca, foi mais preciso do que o *Google*, uma vez que este *site* fez uma previsão de quase 50% e o *Google*, apenas 15,5% de

relevância. Estes resultados podem decorrer da presença dos humanos para proceder à seleção dos documentos que farão parte da base de dados do *AltaVista*, enquanto que, no *Google*, este processo é totalmente automatizado.

4.5 REPRESENTAÇÃO DA SUPERESTRUTURA

A representação da superestrutura foi analisada através da palavra-chave: "História da Paraíba" forma de acesso à informação. Percebemos que, embora sendo uma busca por palavras-chave com termos fechados, entre aspas, procurando dar sentido à frase e não às palavras isoladamente, houve ruídos informacionais no resultado da busca. A recuperabilidade apontou 60% de irrelevância no *Google* e 36,5% no *AltaVista* (Tabela 1). Isto mostra que, no processo de busca, as palavras-chave podem ser extraídas tanto dos títulos do documento, de seus resumos ou ainda do texto integral, utilizando a linguagem natural, sem a consideração do discurso no qual elas estão inseridas, gerando o que Le Guerm (1991) chama de "símbolos sem referência".

Em sistemas de informação tradicional, os resumos são considerados como instrumentos de grande importância para pesquisadores, uma vez que facilitam a seleção do material desejado, ajudando o usuário a decidir sobre a previsão de um determinado item satisfazer suas necessidades, ou seja, contribuindo eficazmente para recuperar a informação. Os resumos das chamadas de cada *homepage* na verdade são apenas extratos dos textos integrais, o que difere de resumo propriamente dito. Segundo Lancaster (1993, p.88), "extrato é uma versão abreviada de um documento que se elabora extraindo frases do próprio documento. [...] o resumo verdadeiro, ainda que inclua palavras do texto, é um texto criado pelo resumidor e não uma criação direta do autor". O autor argumenta, ainda, que as passagens do texto, se bem selecionadas, também podem se tornar pistas importantes para a recuperação de informação, fato constatado neste estudo, onde os resultados das buscas no *AltaVista* alcançaram uma previsibilidade de quase 100% como resposta dos itens relevantes. No caso do *Google*, percebemos que os extratos encontrados nas chamadas não proporcionaram boa

previsibilidade, alcançando apenas 16%, aproximadamente. Desta forma, é válido inferir que este fato ocorre em virtude da ausência ou deficiência de análise do contexto semântico destes dados do extrato, talvez pela inexistência de humanos para a escolha de seus extratos, uma vez que a indexação é totalmente automática.

4.6 COBERTURA

Para legitimar a análise da cobertura em relação a títulos e autores, tomamos como referência a base de dados DEDALUS, apontada pelos professores durante as entrevistas como uma das mais pertinentes para suas pesquisas, levando em consideração a coleção de itens na área de História da Paraíba. Para isto, efetuamos uma busca nesta base, tomando como referência a mesma estratégia de busca utilizada nos *sites Google* e *AltaVista*, quer dizer, adotando-se o sintagma "História da Paraíba", que nos forneceu, inicialmente, uma listagem contendo 116 títulos com os seus respectivos autores. Aqui também, extraímos apenas os resultados expressivos para a investigação, chegando-se assim ao resultado de 90 títulos. De posse desta relação, fizemos o cotejamento destes itens nos documentos das chamadas recuperadas nos *sites Google* e *AltaVista*.

4.6.1 Cotejamento da Cobertura de Conteúdo*: Google e DEDALUS*

Para o cotejamento da cobertura do *Google* e DEDALUS da USP, tomamos por base os 49 itens considerados relevantes no *Google*. No entanto, ao se efetuar o cotejamento, constatamos que um mesmo documento apresentava em diferentes chamadas, então, este total de 49 itens reduziu-se a apenas 32 itens considerados relevantes. Após esta análise, passamos a averiguar a autoria concernente a cada documento, assim, deste total, somente 13 possuíam autoria individual e os 19 restantes não traziam explicitamente a indicação de autoria. Dos 13 docu-

mentos que possuíam autoria, encontramos no *Google* apenas 9 itens em comum com o DEDALUS da USP. No entanto, ao verificarmos os conteúdos destes documentos, constatamos que, na realidade, não se tratava de 9 obras, e sim de única obra, apresentada de várias maneiras. Trata-se da obra intitulada História da Paraíba, cujo autor é o historiador paraibano Horácio Almeida.

Diante do exposto, consideramos como número de item de relevância no cotejamento concernente à cobertura entre o *Google* e DEDALUS da USP apenas 1 (um) item, sobretudo pela questão de conteúdo que o *site* oferece e não pelo número de exemplares de um certo conteúdo. Assim, determinamos como cobertura de título e autor apenas o item em comum com a base de dados USP, alcançando um percentual de cobertura de apenas 1,11 %, conforme Figura 1 a seguir:

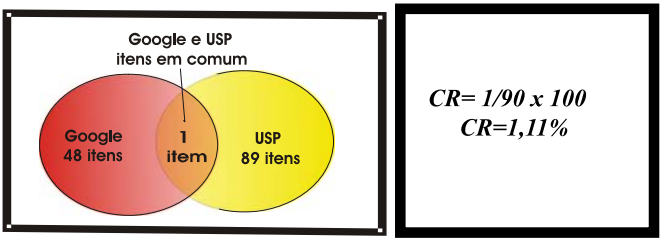

Figura 1: Cobertura Google X DEDALUS
Fonte: Pesquisa direta em outubro de 2003

Embora o grau de duplicidade entre as duas bases tenha se revelado muito baixo (1,11%), os itens de exclusividade do *Google* possuem conteúdos em comum com a base de dados da USP, apenas não possuem a mesma autoria. A lista de conteúdos exclusivos do *site Google*, tomando por base os outros 48 registros que contêm informações predominantes em relação ao assunto, apresentou 09 itens com assuntos periféricos concernentes à História da Paraíba. Estes itens se constituíam basicamente de apontamentos, notas e datas.

Com relação aos 39 itens restantes, pode-se dizer que são documentos que apresentaram conteúdos predominantes sobre o tema História da Paraíba, resultados de pesquisa, artigos, teses e dissertações. Dentre estes conteúdos, encontrou-se elementos que apontam para ou-

tros aspectos sobre a História da Paraíba, destacando-se a História da Policia Militar da Paraíba, Figuras Ilustres do Cenário Paraibano, História de Campina Grande, História dos Teatros Paraibanos, Universidade Eclesiástica na Paraíba, A Mulher Paraibana no Séc. XX, entre outros de considerável importância na construção da memória paraibana.

4.6.2 Cotejamento da Cobertura de conteúdo*: AltaVista e DEDALUS*

No que diz respeito ao cotejamento da cobertura entre o *site AltaVista* e o DEDALUS da USP, consideramos inicialmente os 17 itens que continham informações relevantes em relação ao assunto proposto, História da Paraíba. Entretanto, apenas 13 itens foram considerados relevantes pelas mesmas questões observadas no *site Google*, ou seja, as repetições dos *sites* e dos conteúdos nas páginas de apresentação. Deste total, apenas 7 possuíam autoria e os outros 6 não. Considerando-se os elementos título e autor, encontramos no *AltaVista* apenas 1 (um) item em comum com a base de dados da USP, cujo titulo é "História da Paraíba" e autoria "Almeida, de Horácio", o mesmo registro encontrado no *Google*. Consideramos também como número de item de relevância para análise de cobertura apenas 1 (um) item, ou seja, 1,11% do total de 33 itens, levando-se em consideração o título e o autor. Estes dados coincidem com os mesmos resultados encontrados no *Google*, vede a Figura 2 abaixo.

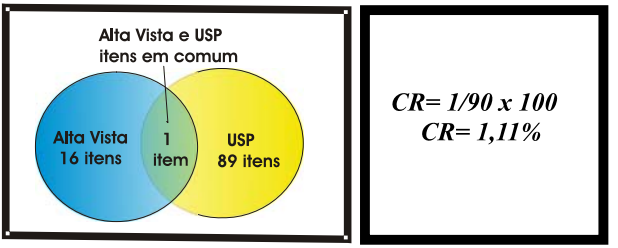

Figura 2: Cobertura *AltaVista* e DEDALUS
Fonte: Pesquisa direta em outubro de 2003

No que diz respeito aos itens de exclusividade do *AltaVista*, a exemplo do *Google*, também verificamos a existência de alguns conteúdos em comum com a base de dados da USP, embora não fosse o mesmo documento e nem tivesse a mesma autoria, ainda assim se mostravam com o mesmo teor de importância no que diz respeito ao conteúdo. Em sua lista de conteúdos exclusivos, e tomando-se por base os outros 16 registros, constatamos que 15 deles apresentavam conteúdos predominantes e apenas 1 (um) portava conteúdo periférico. Portanto, neste *site*, embora tenha recuperado poucos itens, percebemos maior precisão relativamente à informação desejada.

No que diz respeito aos conteúdos predominantes, encontramos elementos que apontam para outros aspectos sobre a História da Paraíba, destacando-se as Figuras Ilustres do Cenário Paraibano, Revolução de 1930, de autoria de José Octávio. Em Torno do Uso Turístico do Patrimônio Histórico: o caso da igreja de Nossa Senhora de Nazaré do Alamagre (município de Cabedelo), de autoria de Carla M. Oliveira, A Universidade Eclesiástica na Paraíba, Paraíba e seus Problemas, História do Jornal "O Norte" no cenário paraibano, A Mulher Paraibana no Séc. XX, entre outros de considerável importância. Como se pode perceber, alguns destes conteúdos são os mesmos encontrados no *site Google*, salientando que não diferenciavam nem mesmo as chamadas das *homepages* (páginas de abertura).

4.6.3 Cotejamento Cobertura *Google* e *AltaVista*

De acordo com os dados analisados, observamos que, apesar do *Google* se apresentar mais expressivo em relação ao *AltaVista*, no que diz respeito ao número de itens recuperados, na cobertura de assuntos, os dois se encontram praticamente no mesmo nível de abrangência, chegando a ter o mesmo item de duplicidade com a base de dados da USP. Isto mostra que nem sempre um grande volume de recuperação de itens necessariamente possibilita a eficácia na recuperação de informação. Entretanto, no ponto de vista do quesito "itens exclusivos", com informações predominantes, as respostas do *AltaVista*

superaram às do *Google*, uma vez que, mesmo tendo recuperado menos itens, apenas 17 contra 49 do *Google*, 16 eram constituídos por conteúdos predominantes em relação à informação desejada. Portanto, levando em consideração este conteúdo nos itens de relevância, a cobertura do *AltaVista* se apresentou como sendo de maior relevância. Este fato pode ser explicado porque o *AltaVista*, como já referimos anteriormente, conta com humanos para efetuar o processamento e a gestão da informação no *site*. Isto colabora para uma representação de informação com maior teor semântico, inclusive vindo ratificar o que está na proposta da *WEB* semântica e nas técnicas da Biblioteconomia e da Ciência da Informação.

4.7 ORDENAÇÃO LÓGICA: chamadas do altavista

Neste quesito verificamos o porquê de, em certas buscas, uma determinada página aparecer em primeiro lugar e em outras, em último. Identificamos a ordem com que cada página foi exposta durante a coleta de dados, juntamente com o seu conteúdo, para, a partir disso, aferir se realmente existe um critério de valor em relação ao conteúdo ou simplesmente há uma troca de ordem para diferenciar os resultados da busca, forjando resultados inéditos em cada navegação realizada. Para tal avaliação, enumeramos todos os itens, visando a uma demarcação de posicionamento durante os dias de coleta, ou seja, cada chamada de *homepage* ganhou um número e permaneceu com ele até o final destas análises.

Em função da exigüidade do tempo e da quantidade enorme de itens, optamos por analisar apenas as 10 primeiras e as 10 últimas chamadas de cada *site*, que constituíam a primeira e a última tela de resultados. Desta forma, foi possível observar, por exemplo, se o primeiro item ou o último mudou de posição ou de colocação durante os oito dias pesquisados. Além disto, foi analisado, também, o conteúdo dos itens desta delimitação para verificar se realmente o conteúdo está sendo considerado ou não, para a posição que ocupa na tela de resultados.

4.7.1 Ordenação lógica*: chamadas do Google*

Através de informações obtidas no próprio *site*, verificamos que o *Google* explica a ordem cronológica de apresentação das chamadas, utilizando como critério de ordenação e colocação o número de vezes que determinada página é consultada. Desta forma, presumimos que as páginas mais consultadas são aquelas que contêm um conteúdo predominante sobre o tema pesquisado, razão pela qual se percebe a necessidade de verificação também dos conteúdos recuperados nessas páginas.

De acordo com os dados coletados, verificamos que, durante todo período de coleta de dados no *Google*, ou seja, durante 8 dias, a ordem dos 10 primeiros itens permaneceu a mesma. Entretanto, com relação aos 10 últimos itens, verificamos algumas alterações de posições. Nos dois primeiros dias de coleta (07 e 08/10), recuperamos na primeira tela 4 itens relevantes e 6 irrelevantes. Deste total, três ocupavam as primeiras posições dos resultados (1º, 2º, 3º), sendo todos de conteúdos predominantes. O último restante dos itens relevantes, em vez de ocupar a quarta posição, tomou a décima posição dessa primeira tela, ou seja, apareceu em último lugar na referida tela, subentendendo-se que os itens concernentes à 4ª e 9ª posições se constituíam justamente dos seis itens irrelevantes.

No terceiro dia de observação, houve pequena alteração de posição do 4º item (irrelevante), intitulado "SBBD/SBDS2000 – História da Paraíba", com o 11º item (relevante), intitulado "Leandro de Lima Lira/Monografia de História da Paraíba" (conforme exemplo abaixo), ficando este último na 4ª posição até o final do período observado.

DE:

SBBD/SBES 2000 - **Historia da Paraiba**
Nos horizontes do passado por Dorgival Terceiro Neto. Sobre o autor: Advogado militante, jornalista, historiador, ex-prefeito de João Pessoa. **...** www.pbnet.com.br/openline/cefet/historia_port.htm - 3k - Em cache - Páginas Semelhantes [Mais resultados de www.pbnet.com.br]

PARA:
Leandro de Lima Lira / Monografia de **História da Paraíba**
Monografia de **História da Paraíba**. ÍNDICE. Capítulo I. 1.1 Antecedentes da conquista da PB. 1.2 A conquista e fundação da PB. 1.3 **...**
www.leandrolimalira.hpg.ig.com.br/historiapb/ - 7k - Em cache - Páginas Semelhantes
[Mais resultados de www.leandrolimalira.hpg.ig.com.br]

Do item 5 até o 7 não houve alteração alguma de colocação de itens relevantes e irrelevantes. Na última tela, observamos que, durante os dias coletados, foram recuperados 5 itens relevantes ao tema, sendo 4 de conteúdo predominante, 1 (um) de conteúdo periférico e 5 irrelevantes ao tema. Nesse caso, observamos que, ainda nas últimas posições, encontram-se itens relevantes para a pesquisa ora proposta, e, se considerar o número de consultas para a ordenação dos itens, uma questão é levantada: será que a mesma prática que acontece na Biblioteca tradicional, onde livros acondicionados nas últimas prateleiras da estante não são consultados pela falta de incentivo por parte dos usuários, já que estão em local desconfortável para a busca, acontece também no ambiente hipermídia? Ou seja, itens que se encontram nas últimas posições, apesar de serem relevantes, não são consultados pela inconveniência da busca. Ora, realizar uma busca para se encontrar informações relevantes em meio a trinta ou até quarenta itens não é uma tarefa tão difícil, mas não deixa de ser cansativa; o que dizer, então, de ter que realizar esta mesma busca numa massa documental de 150 itens, ou até 500, 10.000 itens, como acontece em alguns resultados no ambiente hipermídia? Neste caso, fica quase impossível a recuperação de uma massa considerável de informações relevantes oferecidas por este ambiente, considerando o custo/beneficio em relação ao tempo disponível para tais buscas.

4.7.2 Ordenação das Chamadas*: AltaVista*

Diferentemente do *Google*, este *site*, embora tenha a presença do humano, não apresentou nenhuma informação sobre a existência ou não de critérios de ordenação e apresentação dos resultados de uma

busca, o que dificultou em certo ponto as análises desta categoria. De acordo com os dados coletados, verificamos que, durante os 08 dias de coleta, o *site* permaneceu praticamente com a mesma ordem dos primeiros e dos últimos itens, ou seja, os três primeiros resultados são os relevantes. A única alteração observada nos dez primeiros itens foi uma substituição de um item irrelevante por outro também irrelevante, o que não contribuiu para a melhoria dos resultados. Eis abaixo tal alteração a seguir.

DE

CURSOS DE GRADUACAO DA UFPB
... III Sociologia II (Cultural) Sociologia da Educação I Fundam. Cient. da Comunicação II **História da Paraíba** II **História da Paraíba** I Problemas Sócio-Econ. Contemporâneos Língua Inglesa II Língua ...www.prg.ufpb.br/cursos/1230200C.HTM

PARA
vitrine01
... da Alma Archidy Picado Filho O SEBO CULTURAL - 2001 140 páginas - R$ 15,00 cod. 40120 História da Paraíba **José Octávio de Arruda Mello UFPB - 2000 - 6ª edição 279 páginas - R$ 11,90 - novo cod ...www.osebocultural.com.br/vitrine01.htm**

Em todos os dias de coleta (07 a 14/10) no *AltaVista*, recuperamos, na primeira tela, 6 itens relevantes e 4 irrelevantes. Deste total, três ocupavam as primeiras posições dos resultados (1ª, 2ª, 3ª), sendo todos de conteúdos predominantes. Os três restantes ocupavam as 6ª, 7ª e 8ª posições, ficando a 5ª posição sempre ocupada por item irrelevante. Na última tela de resultados, observamos que, durante todos os dias coletados, foram recuperados 2 itens relevantes, de conteúdos predominantes ao tema, e 8 irrelevantes. Neste caso, observamos, também, como no *Google*, que ainda nas últimas posições se encontraram itens relevantes para a pesquisa. Outro fato observado no *AltaVista*, como também no *Google*, foi a mudança contínua de posições dos itens relevantes e irrelevantes nas outras telas não analisadas em profundidade durante todo o período de investigação. Este fato ratifica mais uma vez a reflexão de Bougnoux (1993) ao abordar a ilusão de se estar bem

informado na sociedade atual, uma vez que as respostas das buscas, aparentemente, se mostraram inéditas a cada navegação, mas, na verdade, ao se analisar tais respostas, percebemos que os *sites* oferecem os mesmos itens em diversas buscas, porém esses itens vêm sempre em nova ordem de posições de apresentação, forjando resultados de ineditismo a cada navegação.

4.7.3 Cotejamento de Ordenação das Chamadas*: Google e AltaVista*

Nas observações efetuadas nos dois *sites*, destacamos o fato de que a ordenação destes caminha, de certa forma, de maneira semelhante, ou seja, os dois priorizam, em suas primeiras telas, as três primeiras colocações com itens relevantes. Já, de maneira geral, ambos os *sites* apresentam itens irrelevantes a partir da quarta posição de apresentação das respostas. Porém, algumas ressalvas devem ser feitas. Por exemplo, na primeira tela de respostas do *Google*, detectou-se a 10ª posição contendo também um item relevante, tanto quanto na última tela consultada. Com relação ao *AltaVista*, além dos itens apresentados nas três primeiras, verificamos que aqueles ocupados alocados no intervalo entre a quinta e a sétima posição se apresentavam também com itens relevantes.

Ressaltamos que, em ambos os *sites*, nas outras telas, havia alterações contínuas de posições de itens, onde, tanto os relevantes como os irrelevantes trocam de posição constantemente. Se tomar por base o critério do *site Google* para ordenação, ou seja, se a determinação das posições, decorre do número de vezes com que cada item é consultado, uma primeira pergunta nos é suscitada: será que os outros itens relevantes encontrados no decorrer da relação não são consultados? Fica difícil, portanto, estabelecer uma resposta em relação à ordenação dos *sites*, uma vez que, da mesma forma que se encontram itens de relevância no início dos resultados, se encontram também no meio da busca e no final, o que nos leva a acreditar em uma ordenação aleatória de itens em ambos os *sites*. Com isto, é lícito assinalar que o número de consultas de

uma determinada página não é indício de sua relevância ao tema que se deseja pesquisar, o que vai de encontro ao critério de relevância para a ordenação dos resultados de uma busca, adotado pelo *Google*.

De posse destes resultados, verificamos que a representação informacional em nível de indexação dos *sites Google* e *AltaVista* se processa de modo diferente. No primeiro, a indexação acontece de forma totalmente automática, comparando-se às buscas efetuadas com a sua base de dados, que é construída a partir da compilação de páginas através da varredura realizada 24 horas, diariamente, em toda a internet, portanto, sem interferência de humanos neste processo. Sendo assim, os resultados de suas respostas se apresentaram sem muita eficácia, embora tenham sido relevantes. Isto pode ser em decorrência do uso de palavras-chave do tipo unitermo, que nem sempre se mostram eficazes para recuperar a informação.

Com relação ao *site AltaVista*, ficou evidente que, embora os resultados tenham sido menores, na realidade a sua eficácia foi bem maior. Isto pode ser em razão da existência de humanos que selecionam e organizam as informações captadas pelos seus motores de busca. Conquanto, também, as palavras-chave do tipo unitermo sejam instrumentos de recuperação neste *site*, como em todos os outros da *WEB*, aqui, tais palavras são analisadas no momento da busca dentro das categorias pré-determinadas, local onde todas as informações estão organizadas e armazenadas em grandes áreas do conhecimento.

5 REFLEXÕES FINAIS

Com a realização deste estudo, verificou-se que os deslumbres de acesso ilimitado à informação propagados pela Internet, foram precoces e imprecisos, em detrimento dos grandes desafios impostos pelo processamento e organização totalmente automáticos da informação, que, sem o auxílio da mão-de-obra dos indexadores, profissionais da informação ou especialistas de determinados domínios, nenhum mecanismo automático consegue sua eficácia plena. Ora, se a recuperação da informação nos serviços tradicionais, com o volume reduzido de do-

cumentos, já exige grandes desafios de processamento e organização, principalmente no que diz respeito à sua representação, quiçá com o explosivo volume existente no ambiente virtual, onde os recursos para a recuperação da informação continuam limitados.

Outra constatação do estudo, ainda na área de processamento, é o fato de que na representação em nível de indexação pouco mudou, em detrimento das grandes mudanças que revolucionaram os suportes e a comunicação do conhecimento humano, pois, mesmo em ambientes hipermídia, tais representações ainda se processam de forma tradicional, utilizando-se de palavras-chave do tipo unitermo que, conforme os resultados e as reflexões de Le Guern (1991), Kuramoto (2002) e Bentes Pinto (2002) já não são mais pistas suficientes para facilitar a recuperação de informação eficaz. Isto ficou patente no estudo, pois desde as primeiras buscas utilizando unitermos como palavras-chave, obtinha-se como respostas todas as páginas que continham as palavras História e Paraíba, sem, no entanto, estabelecer uma relação entre elas, porém, quando se optou pela busca fechada "História da Paraíba", os *sites* a entendiam como um sintagma e forneciam itens com certos conteúdos mais direcionados ao tema.

Os resultados da pesquisa mostram a necessidade de se implementar, ou reformular as políticas de representação em nível de indexação existentes nos *sites*, uma vez que não estão dando conta da demanda informacional de seus usuários, que dispensam muito tempo para encontrar as informações desejadas.

As análises concluíram ainda que o tratamento e a organização, resultantes dos serviços automatizados de indexação dos *sites*, necessitam de reajustes no que diz respeito à função de representar os documentos eletrônicos, tanto do ponto de vista físico (características físicas dos documentos, considerando a especificidade deles: imagem, som, texto) quanto do prisma temático (ou de descrição do conteúdo), para que seu desempenho atinja as expectativas levantadas em torno deste novo canal de informação, até então inalcançadas. Essa inadequação do tratamento da informação reflete concisamente em dois aspectos avaliados: a recuperabilidade e a previsibilidade, ou seja, tanto no *Google* como no *AltaVista*, existe a recuperação da informação, porém, nos dois *sites*, a precisão desta recuperação se revela ausente.

Em relação à previsibilidade, a pesquisa confirmou o que Lancaster (1993) evidenciou claramente sobre a necessidade de maior extensão de representação do conteúdo com vistas a recuperar a informação.

Finalmente, acredita-se que o presente estudo apresenta importante contribuição para a área da Ciência da Informação, fundamentalmente no que concerne à representação de conteúdo em nível de indexação (direcionada a este novo suporte do espaço hipermídia de informação), uma vez que estabeleceu um diagnóstico da organização do conteúdo na Internet através de dois grandes mecanismos da *WEB: Google* e *AltaVista*. Isto se revela de significância fundamental na atualidade, haja vista a Internet estar no centro das discussões atuais sobre o tema proposto, inclusive com a criação da *WEB* semântica que visa oferecer informação com maior significado através da criação de ontologias para a representação indexal no ciberespaço.

Abstract

It presents the result of the exploratory study on the representation of nominal syntagm "História da Paraíba" in the WEB, in the Google and Alta Vista sites, aiming at the investigation of how they proceed the representational organization, in both the context of the indexing, and the covering of contents, considering the headings of the USP database, in the area of "História da Paraíba". The methodology used was the direct search in the Google and in the Alta Vista, as well as the individual intensive observation and non-structural interview, and the content analysis techniques. The results show that the index representation in the Web needs readjustments on how to represent the electronic documents to attend the expectations of this new informational media. Concerning the content covering, it was

observed tha , although it had reached a small percentage of completeness in both Sites, being the USP DEDALUS database, both Sites present the subject of importance in the area of 'História da Paraíba'.

Keywords

INFORMATIONAL REPRESENTATION
INDEXING
AUTOMATIC INFORMATION RETRIVAL

REFERÊNCIAS

AUSTIN, D. PRECIS. ***A manual of concept analysis and subject indexing***. London: Council of the Britsh National Bibliography, 1974.

AZEVEDO NETTO, Carlos Xavier de. ***Signo, sinal, informação: as relações de construção e transferência de significados***. Informação & Sociedade: estudos, v.12, n.2, 2002.

BARDIN, Laurence. ***Análise de conteúdo***. Lisboa: Edições 70, 1977.

BENTES PINTO, Virgínia. Indexação: uma forma de representação do conhecimento registrado. ***Ciência da Informação***. Belo Horizonte, v.22, n.2, p.123 - 134, 2001.

_____. ***A representação do conhecimento através da análise de citações: o caso da UFC***. (encaminhado para ENANCIB 2003).

BOUGNOUX, D. ***Sciences de l'information et de la communication***. Paris: Larousse, 1993.

CHAUMIER, J.; DÉJEAN, M. L'indexation assistée par ordinateur: principes et méthodes. ***Documentaliste-Sciences de l'information***, v.29, n.1, 1992.

FERNANDES, Joliza Chagas. ***Representação da informação na Web:*** História da Paraíba nos *sites AltaVista* e *Google*. João Pessoa: CMCI/ UFPB, 2004. (Dissertação de Mestrado em Ciência da Informação)

GIMENEZ LUGO, Gustavo Alberto; ANDRADE, Marco Túlio Carvalho; SICHMAN, Jaime Simão. ***Recuperação de Informação usando computação nebulosa a partir de documentos com estruturas heterogêneas.*** Disponível em: www.pcs.usp.br . Acesso em: 01 jul. 2002.

GRAND dictionnaire de la psychologie. Paris: Larousse, 1991.

KOBASHI, N. I. ***A Organização e a Transferência de Informação Documentária****:* problemas e perspectivas. A organização do conhecimento e dos sistemas de classificação. Brasília: IBICT, 1996.

KURAMOTO, Hélio. Sintagmas nominais: uma nova proposta para a recuperação de informação. ***DataGramaZero***, v. 1, n. 3, fev. 2002.

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

LAWRENCE S, Giles CL. ***Accessibility of information on the web***. Nature 1999 July, 400:107-9. Disponivel em: http://www.polbr.med.br/arquivo/evba1099.htm.

LARDY, J. P. ***Recherche d'information dans Internet: outils et mèthodes***. Paris: Presses Universitaires de France, 1997.

LE GUERN, Michel. Les ***Descripteurs d'un systéme documentaire***; essai de definition. Actes du colloque traitement automatique des langues naturelles et systémes documentaires, 1984. p. 162-169

LUHN, H.P. ***Keyword-in-context index for technical literature (KWIC index).*** Ney York, *IBM.*1959.

MILSTEAD, J.; FELDMAN, S. ***Metadata*****: cataloging by any other name** Online: the leading magazine for information professionals, v. 23, n. 1, Jan. 1999. Disponível na Internet: <http://www.onlineinc.com/onlinemag/OL1999/milstead1.html>. Acesso em: 25 maio 2000.

PEIRCE, Charles S. ***Semiótica***. São Paulo: Perspectiva, 1977.

SALTON, G. ***The SMART Retrieval System—Experiments in Automatic Document Processing Prentice-Hall.*** Englewood Cliffs, N.J, 1971.

VOLPATO, Gilson Luiz. *Publicação Científica.* Botucatu: Santana, 2002.