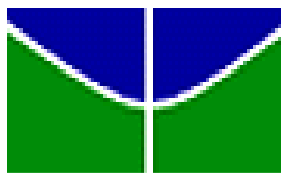


Universidade de Brasília - UnB
Faculdade de Ciência da Informação – FCI
Graduação em Biblioteconomia

**O USO DAS LINGUAGENS NATURAL E
CONTROLADA NA RECUPERAÇÃO DA
INFORMAÇÃO NA WEB:
O CASO DAS LIVRARIAS ELETRÔNICAS**

TAINÁ BATISTA DE ASSIS

Brasília
2010



Universidade de Brasília - UnB
Faculdade de Ciência da Informação – FCI
Graduação em Biblioteconomia

O USO DAS LINGUAGENS NATURAL E CONTROLADA NA RECUPERAÇÃO DA INFORMAÇÃO NA WEB: O CASO DAS LIVRARIAS ELETRÔNICAS

TAINÁ BATISTA DE ASSIS

Monografia apresentada à Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de bacharel em Biblioteconomia.

Professora orientadora: Marisa Bräscher Basílio Medeiros

**Brasília
2010**

A848u Assis, Tainá Batista de

O uso das linguagens natural e controlada na recuperação da informação na web: o caso das livrarias eletrônicas / Tainá Batista de Assis. – Brasília: FCI/UnB, 2010.

90 f. (Monografia de conclusão de curso).

Bibliografia.

1. Recuperação da informação. 2. Linguagem natural. 3. Linguagem controlada. 4. Livraria eletrônica. I. Título.

Aos meus pais,
Maria Lúcia e Djalmir
com muito amor.

AGRADECIMENTOS

Agradeço primeiramente a Deus por me guiar, orientar e por me permitir essa tão sonhada realização pessoal.

Agradeço aos meus pais, Maria Lúcia e Djalmir, pelo incentivo, paciência, apoio e por me aguentarem em todos os momentos de angústia e aflição que ocorreram durante a execução deste trabalho. Também sou eternamente agradecida por se fazerem presentes nos momentos de felicidades e conquistas. Só tenho a dizer que tenho muito orgulho de ser a filha de vocês.

Agradeço a Magda e ao Sylvio por toda contribuição na minha vida e apoio durante o meu estudo. Aos meus irmãos, Paulo, Diogo e Isabela por compreenderem e também me incetivarem para a realização deste trabalho.

A toda a minha família, tias, tios, madrinha, padrinho, primos, primas, vovós, vovôs, pessoas fundamentais na minha formação, o meu muito obrigado.

Ao Roney, meu amor, pelas valiosas contribuições durante todo o meu curso e que soube esperar pacientemente pela conclusão deste trabalho. Agradeço também a sua família por todas as palavras de incentivo.

A minha orientadora, Marisa Bräscher, pela compreensão, tranquilidade e por compartilhar o seu conhecimento para que eu conseguisse concluir este trabalho.

Ao professor Cláudio, por todas sugestões e conversas, e ao Idalécio, por aceitarem fazer parte da minha banca.

Aos amigos que fui formando durante todo o curso, em especial a Fabiane, Yaciara, Wanne, Vivianne, Jonniery, Mariana, Ana Flávia e Raíssa, o meu muito obrigado pela amizade, que tornou os meus dias na faculdade mais alegres.

Aos amigos de cursinho pré-vestibular e da época da escola, agradeço todo o apoio e por compreenderem a minha ausência em alguns momentos.

A todos os meus supervisores e colegas que fiz durante os estágios e que contribuíram para o meu aprendizado.

Por fim, agradeço a todos que direta ou indiretamente contribuíram para a minha chegada até aqui.

“A educação é a arma mais poderosa que você pode usar para mudar o mundo”.

Nelson Mandela

RESUMO

Este trabalho trata da adoção das linguagens natural e controlada para a recuperação da informação disponível na web, em sites de comércio eletrônico. O estudo foi desenvolvido por meio da revisão de literatura do assunto em questão complementada por pesquisa exploratória com trinta usuários que fazem uso frequente ou não da internet. Para o estudo prático, foram previamente definidos dois *sites* de livrarias eletrônicas: a livraria Saraiva e a livraria Cultura e estabelecido previamente um produto comum nos *sites* citados, que corresponde a um livro (Como usar a biblioteca na escola). Os *sites* permitem que os usuários realizem a pesquisa com o uso da linguagem natural, no campo de busca, ou da taxonomia (linguagem controlada). Para a avaliação da qualidade da informação recuperada pelas duas linguagens nesses dois *sites*, foram escolhidos dois critérios: tempo e resultados apresentados. Quanto aos resultados, a busca livre foi considerada pelos entrevistados como a forma de pesquisa que apresentou ser a mais adequada em relação ao tempo e resultados nos dois *sites* de livraria eletrônica, enquanto a taxonomia demandou maior tempo para encontrar o produto. Como considerações finais, observou-se que é de grande importância a utilização das duas linguagens para que atendam de maneira satisfatória a todos as pessoas que realizam buscas nos *sites*.

Palavras-chaves: Recuperação da informação. Linguagem natural. Linguagem controlada. Livraria eletrônica.

ABSTRACT

This work deals with the adoption of natural and controlled languages for information retrieval available on the web, in e-commerce sites. The study was developed through a literature review on the subject supplemented by exploratory research with thirty users who make frequent use of the Internet or not. For the practice study was previously defined two sites of electronic books: the Saraiva Bookstore and Culture and previously established a common product of the sites cited, which is a book (*Como usar a biblioteca na escola - How to use the library at school*). The sites allow users to perform the search with the use of natural language in the search field, or taxonomy (controlled language). To evaluate the quality of information retrieved by the two languages in these two sites were chosen with two criteria: time and results presented. Concerning the results, the search was considered free by respondents as the form of research that had to be adequate in relation to time and results in two sites of electronic bookstore. While the taxonomy has demanded more time to find the product. Lastly, we observed that is very important to use the two languages to meet satisfactorily all the people doing Internet searches.

Keywords: Information retrieval. Natural language. Controlled language. Electronic bookstore.

LISTA DE ILUSTRAÇÕES

Figura 1 – Operador booleano AND	36
Figura 2 – Operador booleano NOT	37
Figura 3 – Operador booleano OR	37
Figura 4 – Interface do <i>site</i> da livraria Saraiva.	42
Figura 5 – Interface da pesquisa com operador booleano E no <i>site</i> da Saraiva.	43
Figura 6 – Interface da pesquisa com operador booleano AND no <i>site</i> da Saraiva.	44
Figura 7 – Livros de Biblioteconomia na livraria Saraiva.....	45
Figura 8 – Interface da pesquisa no <i>site</i> da Saraiva com resultado divergente da anterior.	46
Figura 9 – Interface da pesquisa com a utilização das aspas no <i>site</i> da Saraiva.....	46
Figura 10 – Interface do resultado da busca por assunto no <i>site</i> da Saraiva.....	47
Figura 11 – Interface do resultado da busca no <i>site</i> livraria Saraiva que inclui os livros de Arquivologia.	48
Figura 12 – Interface do resultado da busca no <i>site</i> da Saraiva com os livros de Biblioteconomia na segunda página.....	48
Figura 13 – Interface do <i>site</i> da livraria Cultura.....	50
Figura 14 – Interface da pesquisa com operador booleano AND no <i>site</i> da Cultura.....	51
Figura 15 – Interface da pesquisa com operador booleano E no <i>site</i> da Cultura.....	52
Figura 16 – Interface do resultado da busca com o uso de aspas.	53

LISTA DE QUADROS

Quadro 1 – Indexação pré-coordenada e pós-coordenada.....	21
Quadro 2 – Critérios e características das linguagens natural e controlada	23
Quadro 3 – Operadores booleanos	36
Quadro 4 – Refinamento da busca.....	39

LISTA DE GRÁFICOS

Gráfico 1 – Porcentual de entrevistados de acordo com o sexo	57
Gráfico 2 – Faixa etária entre entrevistados do sexo feminino.....	58
Gráfico 3 – Faixa etária entre entrevistados do sexo masculino	58
Gráfico 4 – Quantidade de entrevistados do sexo feminino quanto à profissão.....	59
Gráfico 5 – Quantidade de entrevistados do sexo masculino quanto à profissão	59
Gráfico 6 – Porcentagem dos entrevistados com relação à frequência do uso da internet	60
Gráfico 7 – Porcentagem dos entrevistados com relação à frequência de pesquisas na internet	60
Gráfico 8 – Porcentagem dos entrevistados em relação ao conhecimento sobre estratégias de busca	61
Gráfico 9 – Quantidade de entrevistados de acordo com a faixa etária que utilizam pelo menos uma estratégia de busca.....	61
Gráfico 10 – Porcentagem dos entrevistados com relação o conhecimento dos <i>sites</i> das livrarias.	62
Gráfico 11 – Porcentagem do total de entrevistados com relação ao caminho utilizado para a busca	62
Gráfico 12 – Grau de satisfação quanto ao resultado da primeira pesquisa	63
Gráfico 13 – Porcentagem dos entrevistados quanto às dificuldades	64
Gráfico 14 – Grau de dificuldades em relação aos itens expostos.....	64
Gráfico 15 – Quantidade de tentativas para a primeira pesquisa.....	65
Gráfico 16 – Grau de diferenças entre os dois modos	65
Gráfico 17 – Diferenças entre os dois modos de busca	66
Gráfico 18 – Quantidade de tentativas para a segunda pesquisa	66
Gráfico 19 – Percentagem dos usuários que utilizaram algum instrumento para a busca.....	67
Gráfico 20 – Grau de satisfação quanto os resultados da pesquisa utilizando instrumentos de busca	67
Gráfico 21 – Caminho utilizado pelos entrevistados para a primeira pesquisa	68
Gráfico 22 – Grau de satisfação na primeira pesquisa.....	68
Gráfico 23 – Dificuldades encontradas no momento da pesquisa.....	69

Gráfico 24 – Quantidade de tentativas para a primeira pesquisa.....	69
Gráfico 25 – Diferenças entre os dois modos de busca no <i>site</i> da Cultura	70
Gráfico 26 – Quantidade de tentativas para a segunda busca no <i>site</i> Cultura.....	70
Gráfico 27 – Uso de estratégias de busca na busca livre	71
Gráfico 28 – Grau de satisfação com uso dos instrumentos de pesquisa	71
Gráfico 29 – Avaliação quanto os critérios tempo e resultados da pesquisa	72
Gráfico 30 – Avaliação dos entrevistados quanto ao tempo da pesquisa	72
Gráfico 31 - Avaliação dos entrevistados quanto aos resultados da pesquisa.....	73
Gráfico 32 - Avaliação dos entrevistados quanto ao tempo da pesquisa	73
Gráfico 33 - Avaliação dos entrevistados quanto aos resultados da pesquisa.....	73

SUMÁRIO

1 INTRODUÇÃO	16
2 PROBLEMA	17
3 OBJETIVOS	17
3.1 Objetivo geral	17
3.2 Objetivos específicos	17
4 JUSTIFICATIVA	18
5 REVISÃO DE LITERATURA	19
5.1 As linguagens natural e controlada	19
5.1.1 Vantagens e desvantagens das linguagens natural e controlada	21
5.2 Ferramentas de representação da informação de linguagem controlada	25
5.2.1 Tesouros	25
5.2.2 Taxonomia	28
5.3 Processamento de Linguagem Natural (PLN)	30
5.3.1 Análise morfológica	32
5.3.2 Análise sintática	32
5.3.3 Análise semântica	33
5.3.4 Análise pragmática	33
5.4 Recuperação da informação	34
5.4.1 Operadores booleanos	36
5.4.2 Operadores de proximidade	37
5.4.3 Truncamento	38
5.5 As livrarias eletrônicas	40
5.5.1 Livraria Saraiva	40
5.5.1.1 Estratégia de busca no <i>site</i> da livraria Saraiva	41
5.5.2 Livraria Cultura	49
5.5.2.1 Estratégia de busca no <i>site</i> da livraria Cultura	49
6 METODOLOGIA	53
6.1 Introdução	53
6.2 A metodologia	54
6.3 Coleta de dados e a população da pesquisa	55
6.4 Questionário	56

6.5 Atividade proposta ao entrevistado	56
7 ANÁLISE DOS DADOS	57
7.1 Perfil do entrevistado	57
7.2 Conhecimento sobre estratégias de busca	60
7.3 Conhecimento dos sites da Livraria Saraiva e Cultura.....	61
7.4 Sobre a Livraria Saraiva.....	62
7.5 Sobre a Livraria Cultura	67
7.6 Busca livre x Busca controlada	71
8 DISCUSSÃO DOS DADOS.....	74
9 CONSIDERAÇÕES FINAIS	77
9.2 Limitações da pesquisa e sugestões para estudos futuros.....	77
REFERÊNCIAS.....	79
ANEXO I.....	84
ANEXO II.....	85
APÊNDICE I.....	86
APÊNDICE II.....	87

1 INTRODUÇÃO

A explosão documental, iniciada por volta dos anos 40, teve como uma consequência, dentre várias outras, a diversidade de suportes documentais possíveis para o armazenamento da informação. Entre esses suportes, o meio digital é o que tem se destacado nos últimos anos pela sua grande capacidade de estocagem de informação de todo o mundo.

Nesse meio digital, a informação tem enfrentado alguns obstáculos em relação à forma de organização na web, e como consequência, esse alto número de informação disponível ao usuário implica no processo de recuperação da informação disponível eletronicamente.

Uma alternativa encontrada para sanar alguns desses bloqueios foi a criação de próprias linguagens, pelos sistemas de recuperação da informação, chamadas de linguagens documentárias (LDs). Estas linguagens têm por objetivo “traduzir” o conteúdo disposto no documento para uma linguagem padronizada. Por outro lado, algumas reclamações são feitas em relação a essas linguagens, como, por exemplo, a queixa quanto à falta de liberdade de expressão dos usuários, dificultando que estes se expressem no momento da busca da forma que pensam.

Ao contrário da linguagem controlada (LC), há a chamada linguagem natural (LN). Essa linguagem se caracteriza por ser a utilizada comumente pelo usuário possibilitando a este mais liberdade no momento da busca, sendo possível utilizar qualquer termo que julga ser necessário para realizar a pesquisa. Porém deve-se ter atenção quanto ao uso, pois é um instrumento no qual as palavras podem apresentar vários significados.

Tendo em vista esses dois instrumentos de representação da informação, este trabalho tem por objetivo fazer uma avaliação das duas linguagens apresentadas, quanto à capacidade de apresentar os melhores resultados na recuperação da informação. Para tanto, serão utilizados para o estudo prático duas livrarias eletrônicas, a Saraiva e a Cultura, que permitem aos usuários em seus *sites* realizarem pesquisas utilizando tanto a linguagem natural quanto a linguagem controlada, por meio da taxonomia.

2 PROBLEMA

A adoção da linguagem natural e/ou da linguagem controlada como ferramenta de representação e organização da informação disponibilizada na web, tem mostrado resultados satisfatórios para os usuários no momento da busca eletrônica?

3 OBJETIVOS

3.1 Objetivo geral

Analisar o uso da linguagem natural e da linguagem controlada na recuperação da informação na web, com foco no usuário final.

3.2 Objetivos específicos

- conceituar as linguagens natural e controlada, bem como apresentar vantagens e desvantagens quanto ao seu uso;
- descrever as características e funções da linguagem documentária, como os tesouros e as taxonomias;
- descrever as características de busca em *sítes* de livrarias eletrônicas;
- reunir, por meio de pesquisa, dados referentes à busca de informação, utilizando a linguagem natural e linguagem documentária;
- relacionar os dados obtidos na pesquisa, apresentando um quadro comparativo entre o uso da linguagem natural e da linguagem controlada.

4 JUSTIFICATIVA

O processo de recuperação da informação disponível eletronicamente ainda é considerado uma tarefa árdua do ponto de vista do usuário final, mesmo com a diversidade de informação disponível na web, bem como com a variedade de tecnologia à disposição para ser utilizada nesse processo.

As dificuldades encontradas nessa etapa de busca por informação estão relacionadas, na grande maioria, pela falta de padronização ou por informações incompletas. Ao realizar uma pesquisa, o usuário final busca a informação com os termos já definidos de acordo com o vocabulário que condiz com o seu contexto. Nem sempre os termos utilizados, porém, são os mesmos que se encontram no sistema de recuperação da informação utilizado para a pesquisa, podendo, portanto, apresentar resultados não satisfatórios ou que não coincidem com o desejado.

A linguagem natural permite maior número de pontos de acesso, e com isso, tem-se um maior número, também, de documentos recuperados. Porém, sem nenhum tratamento há ocorrência frequente de alguns fenômenos linguísticos, tais como a ambiguidade, sinonímia, polissemia entre outros. Na linguagem controlada, esses problemas são evitados, pois os termos utilizados na indexação do documento são resultados de uma padronização do vocabulário, exigindo manutenção periódica do vocabulário.

Com algumas dessas diferenças demonstradas, surge a necessidade de se criarem sistemas que realizem a extração do assunto dos documentos eletrônicos, de forma mais correta, trabalhando tanto com a linguagem natural quanto a linguagem controlada, visando sempre a satisfação do usuário final.

5 REVISÃO DE LITERATURA

5.1 As linguagens natural e controlada

O grande crescimento tecnológico e informacional ocorrido após a Segunda Guerra Mundial, por volta da década de 50, trouxe como consequência a preocupação na forma de organizar a informação e posteriormente no modo de recuperá-la de forma rápida e precisa, atendendo assim à necessidade do usuário. Com a grande variedade de informação, em diferentes linguagens e maneiras distintas de organização e armazenamento desses conhecimentos, se faz necessária a utilização de algum padrão. Dentro desse contexto, surgem as linguagens documentárias. A linguagem documentária (LD) é um instrumento de padronização do vocabulário desses documentos, opondo-se assim, à linguagem natural (LN).

A linguagem natural refere-se à linguagem comum utilizada pelos próprios usuários. As palavras são extraídas diretamente do próprio texto pelo indexador, como Lopes (2002) explica

a linguagem natural (LN) pode ser definida como a linguagem do discurso técnico-científico, e, no contexto da recuperação da informação, Lancaster (2004, p. 200) afirma que "a expressão normalmente se refere às palavras que ocorrem em textos impressos, considerando-se como seu sinônimo a expressão "texto livre". Nas bases de dados, os campos de título e resumo registram os termos da LN (...). (LOPES, 2002, p. 42)

É preciso lembrar que a linguagem natural não apresenta nenhum tratamento, isto é, não há nenhum tipo de controle das palavras para se evitar erros na linguagem. Já a linguagem documentária, também conhecida como linguagem de indexação, é uma linguagem construída com o objetivo de eliminar possíveis problemas do vocabulário natural, como a ambiguidade.

De acordo com Sayão (1996), as linguagens documentárias

são linguagens artificiais geralmente derivadas da linguagem natural. Essas linguagens são chamadas artificiais no sentido em que não resultam de processo evolutivo e necessitam de regras explícitas para seu uso. Normalmente as linguagens documentárias estabelecem uma relação unívoca entre o termo e o conceito, isto é, entre o significante e o significado. Cada termo corresponde a um conceito do sistema de conceitos da área específica com que se está trabalhando. A construção dessas linguagens é um processo complexo e longo. (SAYÃO, 1996).

Segundo Lancaster (2004, p. 19), as linguagens documentárias são vocabulários controlados, sendo basicamente uma lista de termos autorizados estruturada semanticamente. Esta lista deve controlar os sinônimos, diferenciar as palavras homógrafas e fazer uma ligação entre os termos que têm os significados muito próximos entre si. Ainda, de acordo com o autor, são três os principais vocabulários controlados: listas de cabeçalhos de assuntos, os esquemas de classificação bibliográfica, como por exemplo, a Classificação Decimal Universal (CDU) e a Classificação Decimal de Dewey (CDD) e os tesouros.

A linguagem documentária relaciona conceitos limitados em um contexto específico do conhecimento, sendo, portanto uma linguagem construída especificamente para determinado assunto. A linguagem documentária faz o papel de intermediação entre o usuário que busca sanar a sua necessidade de informação e o sistema que utiliza o vocabulário controlado de acordo com a sua área de atuação, buscando traduzir a linguagem utilizada pelo usuário e a adotada no sistema.

De acordo com Cintra (2002, p. 24), “(...) as LDs são, pois, instrumentos intermediários, ou instrumentos de comutação, através dos quais se realiza a “tradução” da síntese dos textos e das perguntas dos usuários”.

Para Gil Urdician (1996) *apud* Boccato (2005)

a linguagem documentária atua em duas fases do processo documentário: no momento da descrição e no da recuperação da informação. O objetivo dessas operações é facilitar a recuperação da informação reduzindo o esforço e o tempo gastos pelo usuário. (GIL URDICIAN, 1996 *apud* BOCCATO, 2005).

As linguagens documentárias podem utilizar três relações distintas: a relação hierárquica, não-hierárquica e a de equivalência. A hierarquização dos termos forma diferentes níveis a partir dos mais gerais aos mais específicos e faz com que a indexação seja mais eficiente e facilite no momento da recuperação da informação.

Quanto à tipologia das linguagens documentárias existem, em relação à coordenação ou combinação dos termos, as pré-coordenadas e as pós-coordenadas. Na pré-coordenada, essa combinação deve ser feita antes da busca da informação pelo usuário, ou seja, na indexação do assunto. Já no caso das pós-coordenadas, a combinação é realizada na hora da busca, podendo-se utilizar como estratégia nessa recuperação, entre outros instrumentos, os operadores booleanos.

Indexação	Vantagens	Desvantagens
Pré-coordenada	<ul style="list-style-type: none"> ■ Termos listados em uma sequência pré-determinada. ■ Evita falsa recuperação. 	<ul style="list-style-type: none"> ■ Representação difícil das várias possibilidades de relações entre os conceitos. ■ Pouca flexibilidade de combinação entre os termos na estratégia de busca. ■ Atualização da terminologia dificultada.
Pós-coordenada	<ul style="list-style-type: none"> ■ Permite diversas combinações de termos. ■ Permite atualização terminológica. 	<ul style="list-style-type: none"> ■ Atribui relevância igual a qualquer termo.

Quadro 1 – Indexação pré-coordenada e pós-coordenada.

Fonte: Apresentação em *slide* de Lopes (2007).

5.1.1 Vantagens e desvantagens das linguagens natural e controlada

Para que haja sucesso na recuperação da informação é necessário que o vocabulário controlado utilizado pelo sistema esteja de acordo com a linguagem natural. Antes da busca pela informação, o usuário deve analisar qual a melhor forma para realizar esse processo tornando-o mais eficiente.

A linguagem natural, como citado anteriormente e diferentemente da controlada, não recebe nenhum tratamento, o que torna a linguagem mais próxima que a utilizada pelos usuários. Por não haver esse processo de “tradução¹”, o custo no sistema é menor, caso que não ocorre com o vocabulário controlado, já que há a necessidade da atualização periódica na linguagem. A partir do quadro 2 a seguir, é possível verificar e comparar as características que apresentam essas duas linguagens.

Crítérios	Linguagem natural	Linguagem controlada
Custos	<ul style="list-style-type: none"> ■ Não há a necessidade de algum treinamento ou profissional, portanto o custo é baixo. 	<ul style="list-style-type: none"> ■ Gasto com a atualização periódica do vocabulário; ■ Necessidade de contratação de profissional para realizar o trabalho.
Revocação e precisão	<ul style="list-style-type: none"> ■ As palavras são extraídas do próprio texto e, portanto, há possibilidade de maior número de pontos de acesso. ■ Maior revocação de documentos e menor a precisão. 	<ul style="list-style-type: none"> ■ O número de pontos de acesso é reduzido. ■ Maior precisão na recuperação da informação.
Controle	<ul style="list-style-type: none"> ■ Não há controle das palavras, podendo haver: <ul style="list-style-type: none"> • Ambiguidade; • Sinonímia; • Homonímia. 	<ul style="list-style-type: none"> ■ Há o controle das palavras utilizadas por meio da padronização do vocabulário.
Flexibilidade	<ul style="list-style-type: none"> ■ Totalmente flexível. ■ Liberdade de expressão na busca. 	<ul style="list-style-type: none"> ■ Não-flexível, rígido. ■ Restringe a forma de expressão dos usuários.
Tempo	<ul style="list-style-type: none"> ■ O tempo gasto na busca é maior. 	<ul style="list-style-type: none"> ■ O tempo de busca é menor, com o uso de termos controlados.
Esforço	<ul style="list-style-type: none"> ■ Maior esforço intelectual, já que o próprio usuário deverá pensar nos termos de 	<ul style="list-style-type: none"> ■ Menor esforço, pois os termos já estão sugeridos para os usuários.

¹ Tradução: Segundo Lancaster (2004, p.15) tradução “envolve uma decisão sobre quais dos rótulos disponíveis melhor representam X, Y e Z”.

	busca.	
Compatibilidade	■ Compatível com qualquer sistema.	■ Cada sistema pode ter sua própria linguagem, o que dificulta a compatibilidade.
Linguagem	■ Utiliza a linguagem comum do usuário.	■ Tende se afastar da linguagem utilizada pelo usuário.
Facilidade	■ Maior assimilação pelo usuário e com isso é de fácil manuseio.	■ Dificuldade no uso pelo usuário.

Quadro 2 – Critérios e características das linguagens natural e controlada.

Fonte: Elaboração própria.

Essas características apresentadas são importantes para serem analisadas e avaliadas. Os custos estão relacionados aos gastos financeiros, tais como a necessidade de haver um treinamento periódico do profissional ou a contratação de novos profissionais e a manutenção do vocabulário controlado, influenciando também no controle do vocabulário. O tempo gasto na pesquisa e o esforço estão relacionados diretamente ao usuário e como este avalia o sistema de recuperação da informação. A linguagem influencia o usuário quando este faz a avaliação da facilidade e flexibilidade no manuseio com o sistema, pois há a possibilidade de utilizar a sua própria linguagem ou buscar entender a linguagem utilizada pelo sistema. A compatibilidade relaciona-se com a linguagem adotada pelo sistema, pois a linguagem natural pode ser inserida em qualquer sistema, já a linguagem controlada deve seguir a que o sistema usa, fazendo assim, que vários sistemas contenham diferentes linguagens e compatibilidades. A revocação e precisão estão ligadas à qualidade da informação recuperada.

A revocação é o número de documentos relevantes à pesquisa recuperados dentre o total de documentos relevantes existentes no sistema de informação. Assim, para que uma busca seja considerada com um número elevado de revocação é necessário que a maioria dos itens pertinentes à pesquisa sejam recuperados. A precisão, entretanto, leva em consideração a competência de se recuperar apenas os documentos sobre o assunto pesquisado e é medida pela relação entre o número de documentos relevantes recuperados e o número total de documentos recuperados.

As características apresentadas pelas duas linguagens devem ser levadas em consideração no momento da busca. Knapp (1982) *apud* Lopes (2002, p. 48) enumera os usos da linguagem natural na busca para se ter maior eficiência:

para tópicos específicos; para temas atuais; para novas terminologias ainda não incluídas nas LC; para uma busca retrospectiva em que o conceito da LC é muito recente e não cobre os anos anteriores; quando o termo da LC é muito abrangente ou muito específico; para pesquisa em várias bases de dados; para identificação imediata de palavras de títulos dos documentos; para complementação de citação bibliográfica incompleta. (KNAPP, 1982 *apud* LOPES, 2002, p. 48).

Ainda, para ter maior aproveitamento, faz-se necessária a utilização de alguns recursos para a busca utilizando a linguagem natural, tais como a truncagem e o uso de operadores de adjacência ou de proximidade. Na truncagem, a técnica é utilizar a raiz da palavra, mas sem apontar as derivações dessa palavra, sejam sufixos ou prefixos, como por exemplo: pedagog* - pedagogia, pedagogo. Já o uso dos operadores “permite especificar, na estratégia, a posição relativa de dois ou mais termos entre eles próprios” (LOPES, 2002).

Para que o resultado de uma busca seja mais eficaz, vários autores afirmam que é necessário haver a combinação entre a linguagem natural e controlada, como afirma Muddamalle (1998, p. 887, tradução nossa ²), “(...) linguagem natural e a linguagem controlada não podem mais ser tratadas como técnicas separadas, mas devem sempre ser tratadas em conjunto como uma combinação ideal”.

Porém, é importante que o próprio usuário avalie anteriormente a ferramenta a ser utilizada, para que se tenha um melhor desempenho na recuperação da informação.

² Texto original: (...) natural language or controlled vocabulary can no longer be treated as separate techniques, but should always be treated together as an “ideal” combination.

5.2 Ferramentas de representação da informação de linguagem controlada

5.2.1 Tesouros

O termo *tesauro* tem origem do grego (*thesaurós*) e do latim (*thesauru*) o qual significa tesouro, repositório, e ainda para alguns autores tesouro de palavras. Inicialmente era considerado como sinônimo para dicionário, mas o termo teve destaque e um novo conceito a partir da obra de Peter Mark Roget, *Thesaurus of English Words and Phrases*, publicada em Londres no ano de 1852. Diferentemente de um dicionário, no qual os elementos seguem organizados em forma alfabética, nesta publicação os termos apareciam na ordem de acordo com as ideias ou significados que representavam, criando-se assim, relacionamentos entre os conceitos destes termos a partir dos vários significados que apresentavam.

Esse termo começou a ser adotado por volta de 1940 na área da Ciência da Informação, como uma ferramenta no processo de organização e recuperação da informação. A partir do trabalho de Hans Peter Luhn, na década de 1950, que utilizou o termo *Tesauro* para nomear o seu sistema de palavras autorizadas, define-se um novo instrumento da linguagem documentária.

No ano de 1977, a UNESCO publicou a sua primeira edição do Manual de Construção de Tesouros, o qual inclui, atualmente, termos para as áreas de educação, ciência, política, direito, economia, cultura, ciências sociais e humanas, de informação e comunicação, e outros termos ligados aos países. Ainda, a UNESCO define tesauro como “vocabulário controlado e dinâmico de termos relacionados semântica e genericamente, que cobre de forma extensiva um campo específico do conhecimento” (DODEBEI, 2002, p. 67).

Segundo Gomes (1996), “tesauro é um conjunto de termos semântica e genericamente relacionados, cobrindo uma área específica do conhecimento. É um instrumento da indexação/recuperação da informação”.

A UNESCO *apud* Campos (2001), define tesauro segundo a função como

dispositivo de controle terminológico usado na tradução da linguagem natural dos documentos, dos indexadores ou dos usuários numa linguagem do sistema (linguagem de documentação, linguagem de informação) mais restrita. (UNESCO *apud* CAMPOS, 2001).

De acordo com Cavalcanti (1978, p. 27)

tesauro é uma lista estruturada de termos associados empregada por analistas de informação e indexadores, para descrever um documento com a desejada especificidade, em nível de entrada, e para permitir aos pesquisadores a recuperação da informação que procura. (CAVALCANTI, 1978, p.27).

A ideia essencial de um tesauro é que esse instrumento flexível possa, de forma fácil e precisa, organizar e estruturar a linguagem natural, objetivando localizar em tempo hábil os documentos, sanando assim, a necessidade de informação dos usuários.

Por meio de um relacionamento lógico e hierárquico, sua principal função é servir de instrumento capaz de controlar o vocabulário, utilizando assim somente termos autorizados para determinado conhecimento, evitando-se problemas linguísticos, tais como sinonímia, polissemia e homonímia³. Esses termos são chamados descritores. Ainda, o tesauro é considerado um sistema pós-coordenado, formado por meio das relações entre os conceitos.

Quanto maior for o número de termos utilizados por um indexador na construção de um tesauro, maior serão os pontos de acesso do documento, sendo portanto mais facilmente recuperado. Em outras palavras, quanto maior a exaustividade da indexação maior é a revocação.

Segundo Carlan (2008, p. 42), pode-se dividir a estrutura de um tesauro em duas partes: a base teórica e a base técnico-operacional. Na base teórica, quatro aspectos são considerados mais relevantes:

- a ideia de conceito como unidades de conhecimento;
- termo que seria a designação do conceito e indivisível;
- as propriedades gerais dos conceitos denominadas categorias e;

³ Sinonímia: Segundo Cintra (2002), “sinonímia é uma relação de equivalência entre, ao menos, duas palavras. Através dela não se afirma a identidade entre os elementos envolvidos na relação”.

Polissemia: São os vários significados que uma palavra assume.

Homonímia: Palavras que tem a mesma grafia ou pronúncia, mas com significados diferentes.

- as facetas que são as características do grupo no qual o conceito está inserido.

Ainda de acordo com a autora, a segunda parte, a base técnico-operacional, seria o processo de construção do tesauro, também dividida em quatro etapas:

- o planejamento, onde se delimita a terminologia do tesauro e demais fatores essenciais;
- a coleta de termos, o controle terminológico com o objetivo de evitar erros da linguagem natural, como a ambiguidade e sinonímia;
- estabelecimento das relações entre os conceitos e;
- modo de apresentação desse tesauro.

Para que os conceitos sejam estruturados dentro de um contexto é preciso que este também esteja explicitado no tesauro. Esse contexto é dado pelas Notas Explicativas, representado por NE. É nesse espaço, portanto, que se conceitua o termo adotado.

O estudo das relações conceituais, ainda, é necessário para que se estabeleça dentro do tesauro uma ordenação e entendimento dos termos utilizados no tesauro. Assim temos as relações termológicas:

- relação de equivalência: ocorre geralmente entre os termos sinônimos ou quase-sinônimos, simbolizada por USE – *used* e UP – usado para. Segundo Dodebei (2002) “quando um conceito pode ser representado por vários símbolos distintos ou quando se quer reduzir, por questões pragmáticas, os níveis de implicação conceitual, estabelecem-se as relações de equivalência semântica”.

Exemplo: Sinônimo

Motorista de caminhão

Caminhoneiro

USE: Caminhoneiro

UP: Motorista de caminhão

- relações hierárquicas: apresentam a subordinação entre os conceitos, utilizando termos subordinados ou específicos – TE e termos superordenados ou gerais – TG. Como define Dodebei (2002) “se dois conceitos diferentes possuem

características idênticas e um deles possui uma característica a mais do que o outro, então entre eles se estabelece a relação hierárquica ou relação de gênero-espécie”.

Exemplo: Fruta cítrica

Limão

TE: Limão

TG: Fruta cítrica

- relações associativas: são os termos que não apresentam nem relação de equivalência nem hierárquica, mas é possível fazer a relação mentalmente. Nesse caso, utiliza-se o símbolo TR. Exemplo: Agroindústria

TR: Economia agrícola

- relações partitivas: usada para relacionar o conceito entre um todo e suas partes, ou entre as partes que se compõem um todo. As suas partes são representadas pelo Termo Específico Partitivo (TEP) e o termo genérico pelo Termo Genérico Partitivo (TGP). Exemplo de Carlan (2008):

Sistema Nervoso Central

TGP: Sistema Nervoso

TEP: Cérebro

TEP: Medula

Como exposto anteriormente, a elaboração de um tesouro requer tempo, conhecimento e pessoas especializadas para trabalhar nesse processo, pois cada tesouro construído é único dentro da área de conhecimento. Um tesouro exige manutenção periódica, pois novos termos são criados quase que diariamente e este precisa estar atualizado para que continue sendo um instrumento eficiente na recuperação da informação.

5.2.2 Taxonomia

A palavra taxonomia, também escrita por alguns autores como taxionomia, tem origem do grego, significando *táxis* – grupo e *nomos* – lei, ciência. O termo é definido como um instrumento para classificar de forma hierárquica os objetos. Um exemplo simples e pioneiro de taxonomia é a de Karl von Linné, que em 1735

divulgou o trabalho, a Classificação Geral dos Seres Vivos, no qual estavam classificados os seres vivos em grupos de acordo com as suas características em comum.

Segundo Terra et al. (2005, p.1), a taxonomia é

um sistema para classificar e facilitar o acesso à informação, e que tem como objetivos: representar conceitos através de termos; agilizar a comunicação entre especialistas e entre especialistas e outros públicos; encontrar o consenso; propor formas de controle da diversidade de significação; e oferecer um mapa de área que servirá como guia em processos de conhecimento. (TERRA et al., 2005, p.1).

Apesar de ser um instrumento já consolidado, a taxonomia vem sendo crescentemente utilizada no meio empresarial, como um instrumento estratégico para os negócios e um mecanismo importante na elaboração de sites corporativos. Esse fato se dá por causa do aumento do volume de informações nas últimas décadas, o que leva as empresas utilizarem e organizarem a informação para servir como um ponto forte e distinto das demais organizações.

Segundo Terra et al. (2005, p. 2)

no contexto do mundo de negócios, podemos definir de maneira simplificada que taxonomias são regras de alto nível para organizar e classificar informação e conhecimento. E isto é essencial, pois organizações que não conseguem fazer isto acabam não compreendendo as atividades agregadoras de valor dentro delas mesmas; atividades estas que crescentemente estão baseadas na produção e uso de informação e conhecimento. (TERRA et al., 2005, p. 2).

Ainda, de acordo com Vogel (2009, p. 1)

(...) sabemos que taxonomia é um sistema para classificar informações – seus assuntos, naturezas, tipos, etc. O que significa que ela é a lista controlada de palavras-chave que irá rotular a informação que a empresa produz. (VOGEL, 2009, p. 1).

As taxonomias elaboradas pelas empresas devem estar de acordo com o conhecimento que esta aborda e principalmente, devem fazer parte do conhecimento de seu cliente, com isso facilitando o processo de recuperação da informação. Essa ferramenta terá como papel na instituição, organizar e controlar os fluxos informacionais de maneira que toda informação seja tratada e padronizada.

As taxonomias não são iguais. Cada organização elaborará uma taxonomia diferente, já que os contextos são distintos. Mas o objetivo deve ser o mesmo: facilitar para que as informações sejam encontradas em curto tempo por meio de categorias bem estruturadas. Deve, também, permitir que a pesquisa pela informação possa ser feita por diversos campos, como autor, tipo de formato do arquivo, título etc.

Terra et al (2005) definem alguns critérios que devem ser seguidos para a elaboração de uma taxonomia, são eles:

- comunicabilidade: a linguagem da taxonomia deve ser clara e seus termos devem apresentar os conceitos mais objetivos para o público.
- utilidade: a taxonomia deve conter apenas aqueles termos que são utilizados pela instituição, não sendo necessário descrever termos que não fazem parte do cotidiano desta.
- compatibilidade: “a taxonomia deve conter somente estruturas do campo que se está ordenando e que façam parte das atividades ou funções da organização”.
- estimulação: uma taxonomia deve estimular o usuário a seguir com a busca pela ferramenta. Este último critério está diretamente ligado a linguagem utilizada pela taxonomia, que deve ser clara, tornando-a atrativa.

Todos esses critérios visam à elaboração de uma taxonomia eficiente para a organização. Vale lembrar que não somente aqueles que fazem parte do processo de construção de uma taxonomia devem estar cientes dos termos adotados e as categorias sugeridas, mas sim todos que fazem parte da empresa devem entender o universo da taxonomia.

5.3 Processamento de Linguagem Natural (PLN)

Quando um sistema de recuperação da informação utiliza a linguagem natural é necessário aplicar métodos do chamado Processamento de Linguagem Natural (PLN) com o objetivo de alcançar maior precisão da informação recuperada.

Segundo Oliveira (2009, p. 1)

a tarefa de processar uma linguagem natural permite que os seres humanos comuniquem-se com os computadores da forma mais "natural" possível, utilizando a linguagem com a qual mais estão habituados. Elimina-se, desta maneira, a necessidade de adaptação a formas inusitadas de interação, ou mesmo o aprendizado de uma linguagem artificial (...). (OLIVEIRA, 2009, p. 1).

Os primeiros estudos nessa área começaram na década de 50 com as tentativas de informatizar a linguagem natural. A partir da década de 60, alguns computadores eram capazes, ainda de forma básica, compreender e responder questões em um idioma, o inglês. Nessa época, o programa que teve mais destaque foi o Eliza, desenvolvido por Joseph Weizenbaum, o qual se passava por um psiquiatra que dialogava com usuários comuns, para que com isso, um programa que utilizasse o PLN fosse colocado em prática e passível de testes. O programa se passava tão verdadeiro aos pacientes, que alguns desses usuários, solicitavam que o criador, Weizenbaum, se retirasse da sala para que ficassem a sós com a máquina.

O processamento da linguagem natural também é aplicável em outras áreas, e não apenas nos sistemas de recuperação da informação, como nas ferramentas de acesso aos bancos de dados, na elaboração de resumos e na tradução automática. Há duas abordagens comumente aplicáveis ao uso do PLN: a abordagem estatística e a linguística.

O PLN baseado em conceitos estatísticos para realizar a análise de textos completos começou a ser estudado e incentivado em razão do grande volume de informação em meio digital. De acordo com Gasperin e Lima (2001, p. 3)

o processamento estatístico da linguagem natural consiste no uso de abordagens quantitativas para o processamento automático de textos. Uma coleta de informações é realizada sobre uma grande base de textos, denominada corpus, para o levantamento das probabilidades de ocorrência de palavras, seqüências de palavras, categorias de palavras, etc (GASPERIN e LIMA, 2001, p. 3).

A abordagem estatística, portanto, está relacionada diretamente com os dados referentes, como por exemplo, à frequência de palavras em textos.

A abordagem linguística envolve algum nível de análise da linguagem. Para que o computador processe uma sentença em linguagem natural é importante que

as informações relativas à morfologia, sintaxe, semântica e pragmática sejam mantidas, como descritas a seguir.

5.3.1 Análise morfológica

A análise morfológica consiste em separar as palavras em unidades isoladas em uma sentença. Os elementos mórficos analisados são: a raiz da palavra, os afixos (prefixos e sufixos), vogal temática e a vogal e consoante de ligação. Segundo Oliveira (2009, p. 3) o analisador morfológico identifica palavras ou expressões isoladas em uma sentença, sendo este processo auxiliado por delimitadores (pontuação e espaços em branco). Esse processo de análise morfológica é fundamental, pois o sistema que compreende o significado de cada palavra em uma frase será capaz posteriormente criar sentenças com estruturas coerentes.

5.3.2 Análise sintática

De acordo com Cegalla (2005, p. 319) a análise sintática “examina a estrutura do período, divide e classifica as orações que o constituem e reconhece a função sintática dos termos de cada oração”. A separação de cada termo permite a análise dos sintagmas, que são a unidade de uma ou mais palavras que representam uma função na oração. Como explica Oliveira (2009, p. 4)

a análise sintática de uma oração em português deve levar em conta os seguintes sintagmas: termos essenciais (sujeito e predicado), termos integrantes (complementos verbal e nominal) e termos acessórios (adjunto adverbial, adjunto adnominal e aposto). A análise do período, por sua vez, deve considerar o tipo de período (simples ou composto), sua composição (por subordinação, por coordenação) e a classificação das orações (absoluta, principal, coordenada ou subordinada). (OLIVEIRA, 2009, p. 4)

Na área da computação, a etapa da análise sintática é conhecida por *parsing* na qual a sentença é estruturada de forma hierárquica, na qual cada palavra tenha agregada a sua importância sintática.

5.3.3 Análise semântica

A análise semântica consiste em compreender os significados das palavras. Essa etapa envolve diferentes situações, tais como a ambiguidade e a diferenciação entre os significados e sentidos que as palavras apresentam. A função é analisar o sentido das estruturas das palavras que foram reagrupadas pelo analisador sintático, uma vez que o analisador morfológico permitiu identificar estas palavras individualmente (Oliveira, 2009, p.4).

5.3.4 Análise pragmática

A análise pragmática permite que o significado da sentença, agora já estudado, seja aplicado ao contexto em que está inserida. É a etapa em que o todo é interpretado e não mais apenas as partes que o compõe.

Segundo Morellato (2008)

enquanto acha-se que o futuro das máquinas de busca caminha para a utilização de pesquisas em linguagem natural, o Google acha que não. Segundo a maior empresa de buscas do mundo, não existe tanta diferença em perguntar "Qual o homem mais rico do mundo?" ao invés de "homem mais rico mundo". Além disso, eles acreditam que os usuários não vão querer digitar perguntas, já que, segundo pesquisas, cerca de 50% das consultas feitas ao site envolvem uma ou duas palavras (MORELLATO, 2008).

Os métodos de busca empregados nas livrarias no comércio eletrônico são normalmente modelos de PLN estatísticos e não linguísticos. Por esta razão podem apresentar problemas relativos ao entendimento da pergunta elaborada pelo usuário no momento da busca por algum produto.

5.4 Recuperação da informação

O grande volume de informação acessível ao usuário na *world wide web* torna o processo de recuperação de informações mais específicas muitas vezes trabalhoso. Mesmo com tanta informação disponível, não será útil para aqueles que pesquisam e não as encontram. Porém, esse não é o único fator que dificulta a busca por informações.

Na web, além do volume, dos diferentes idiomas, dialetos e culturas, outros fatores dificultam a recuperação de informação: limitações dos sistemas de busca, confiabilidade das informações localizadas, volatilidade dos documentos, velocidade das conexões e incompatibilidade tecnológica entre o microcomputador do usuário e os recursos utilizados pelos provedores de conteúdo da web. (YAMAOKA, 2003, p. 1).

A preocupação com o desenvolvimento de sistemas de recuperação eficientes justifica-se, já que pesquisas demonstram que a insatisfação do usuário no momento da busca é grande, com quase 70% de frustração (Roper Strach, 2000 apud Nachmias, 2002, p. 4). Ainda, outra pesquisa demonstrou que o tempo gasto pela metade dos usuários que pesquisam online é igual ou superior a 70% (Berrier, 2000 apud Nachmias, 2002, p. 3).

De acordo com Branski (2004, p. 71), o usuário precisa ter um conhecimento das ferramentas de busca disponíveis para que suas pesquisas sejam eficientes, como todas as formas possíveis de comunicação e as linguagens de busca.

Para Nachmias (2002), as competências necessárias para um usuário realizar uma busca com resultados satisfatórios é este ter a habilidade de “aplicar as regras da lógica booleana⁴ (por exemplo, e, ou), a compreensão de como a informação está organizada, pensamento crítico que permitam escolhas eficazes e do conhecimento de notações da Internet”.

Há uma diferença no que diz respeito ao processo de recuperação da informação e os sistemas de recuperação da informação. Segundo Araújo Jr. (2007, p. 72)

⁴ Lógica booleana: vide página 36.

a recuperação da informação é reconhecida como a recuperação de referências de documentos em resposta às solicitações (demandas expressas por informação). Já os sistemas de recuperação da informação dizem respeito a um sistema de operações interligadas para identificar, dentre um grande conjunto de informações (uma base de dados, por exemplo), aquelas que são de fato úteis, ou seja, que estão de acordo com a demanda expressa pelo usuário. (ARAÚJO JR., 2007, p. 72).

Segundo Lopes (2002, p. 60)

esses sistemas possibilitam o planejamento de estratégias de busca com maior nível de complexidade envolvendo vários conceitos na mesma estratégia; permitem a utilização de busca de palavras apenas dos títulos e resumos dos documentos, isto é, termos da linguagem natural; buscam os termos específicos de linguagens controladas, nos campos de descritor; buscam por autores; por ano de publicação; por títulos de periódicos; por classificação; permitem, também, a busca de conceitos compostos ou simples e a possibilidade de truncagem de raízes de palavras e de substituição de caracteres no meio dos termos, dentre outros recursos de recuperação. (LOPES, 2002, p. 60).

Portanto, o principal objetivo desses sistemas de recuperação é levar a informação desejada ao usuário na tentativa de eliminar assim, a sua necessidade daquela informação. O uso de uma estratégia de busca utilizada pelo usuário pode ser definida como a reunião de operações e tentativas necessárias para que a informação seja encontrada de forma eficaz.

Para facilitar a pesquisa na *web* geral, alguns sistemas de buscas se encontram disponíveis aos usuários, tais como os diretórios por assuntos e os mecanismos de busca, também chamados de *search engines*. Os primeiros surgiram como uma primeira tentativa de facilitar a localização da informação por meio de uma organização hierárquica. Como exemplo, tem-se o diretório: *Yahoo!*⁵. Já os mecanismos de busca “são programas que tem três funções básicas – identificar páginas da *web*, indexar estas páginas em um banco de dados e um mecanismo de pesquisa como interface” (Yamaoka, 2003, p. 5). A recuperação nesse tipo de instrumento é realizada pelo critério de relevância. O exemplo mais expressivo é o *Google*⁶.

⁵ Site do Yahoo! no Brasil: www.yahoo.com.br

⁶ Site do Google no Brasil: www.google.com.br

De maneira geral, os recursos de busca oferecem aos usuários diferentes instrumentos com o objetivo de recuperar rapidamente a informação. Essas outras opções de pesquisa são descritas a seguir neste trabalho.

5.4.1 Operadores booleanos

Os operadores booleanos foram criados pelo matemático inglês George Boole por volta do século XIX, no qual a ideia é usar termos monossilábicos ou expressões simples para serem utilizados na recuperação da informação. Esses operadores permitem ao usuário fazer a combinação das palavras-chaves da informação desejada. São três os operadores booleanos utilizados pelos sistemas de buscas: E (AND), OU (OR) e NÃO (NOT).

Português	Inglês	Sinais
E	AND	+
NÃO	NOT	-
OU	OR	!

Quadro 3 – Operadores booleanos

O operador E, no inglês AND, tem como função restringir a pesquisa, recuperando apenas os documentos indexados que contenham os termos utilizados para a pesquisa, ou seja, só irão ser apresentadas ao usuário as páginas que contenham os dois termos. O resultado desta pesquisa é a intersecção dos documentos indexados pelos dois termos utilizados.

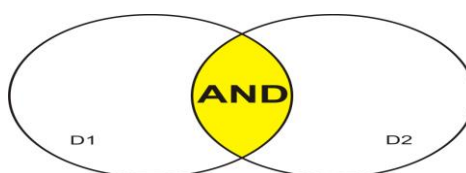


Figura 1 – Operador booleano AND

O operador NÃO (NOT no inglês) faz com que o primeiro termo seja incluído no resultado da pesquisa e o segundo termo seja excluído, sendo, portanto um

operador da diferença entre os documentos. Esse operador deve ser utilizado com maior atenção do usuário, pois informações relevantes poderão não ser recuperadas.

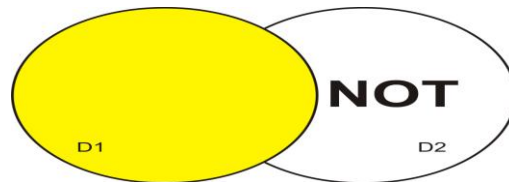


Figura 2 – Operador booleano NOT

Por fim, o operador OU (OR) faz com que o resultado da pesquisa seja maior, já que esse operador permite que sejam recuperados os documentos que contenham os dois termos, um termo ou só o outro termo. Essa combinação de termos é equivalente a união entre os documentos.

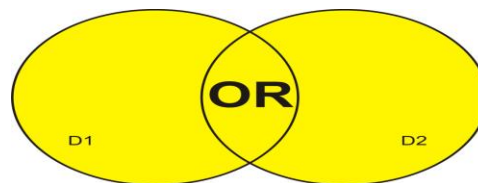


Figura 3 – Operador booleano OR

Utilizar essas possíveis combinações dos operadores booleanos faz com que a busca do usuário se torne mais restritiva. O uso desses operadores lógicos se torna importante também para que possível avaliação da qualidade e relevância da informação recuperada.

5.4.2 Operadores de proximidade

Os sistemas de recuperação da informação, atualmente, dispõem de recursos que recuperam os documentos por meio da relação de proximidade entre as palavras do texto. Os operadores utilizados para exercer essa função de proximidade são o NEAR e o ADJ, também chamados de operadores de texto.

Ao utilizar o operador NEAR, os termos definidos pelo usuário para a busca devem encontrar-se geralmente na mesma frase do texto. O NEAR tem como função recuperar o documento no qual os termos estabelecem uma relação conceitual. Esse fato é que faz a diferença entre esse operador e o AND, o qual busca por documentos que contenham os termos em qualquer parte do texto.

O operador ADJ, *adjacent*, é utilizado para determinar a ordem com que os termos devem se apresentar no texto. Portanto, se o usuário realiza uma pesquisa utilizando o termo1 adj termo2 terá um resultado diferente se pesquisar por termo2 adj termo1.

5.4.3 Truncamento

O processo de truncamento consiste em usar parte da palavra, podendo ser o início, o meio ou o fim, mas geralmente, usando o radical⁷. A busca terá como resultado a parte da palavra selecionada, por exemplo, *administra** que tem como resultado as variações *administrativo*, *administração*, *administrar*. Normalmente, o símbolo utilizado é o asterisco (*).

Há ainda outras técnicas possíveis que os usuários podem utilizar para refinar a sua busca. Essas operações encontram-se resumidas e explicadas no quadro de Branski (2004, p. 76).

Objetivo	Comandos	Função	Exemplos
	Nenhum	Localiza páginas que contenham todos os termos pesquisados, qualquer que seja a ordem em que são apresentados	Comércio exterior Retornará documentos onde constem as duas palavras, qualquer que seja a ordem
Incluir	+ and e	Localiza páginas que contenham todos os termos pesquisados, qualquer que seja a ordem em que são apresentados	+ comércio + trigo comércio AND trigo Retornará documentos onde constem as duas palavras, qualquer que seja a ordem
Excluir	+ - not	Exclui páginas que contenham o termo	+planeta-casseta planeta NOT casseta

⁷ Radical: Parte invariável de um vocábulo. Não apresenta prefixo ou sufixo.

	and not	selecionado	planeta AND NOT casseta Retornará páginas que contenham a primeira palavra, excluindo as que contenham a segunda.
Frase exata	aspas	Localiza páginas que contenham a frase exata	“comércio internacional” retornará documentos onde constem as duas palavras na ordem exata
OR OU	Buscar qualquer termo	Localiza páginas que contenham qualquer um dos termos pesquisados	feminina OR feminilidade retornará documentos onde constem qualquer uma das duas palavras
Proximidade	NEAR	Define quão próximo os termos devem aparecer	lua NEAR rio retornará páginas que contenham as duas palavras separadas por uma certa distância, definida pelo pesquisador
Buscar <i>sites</i>	host:site:domain	Localiza páginas num computador específico	host:nasa.gov venus localiza, somente no site da NASA, páginas que contenham a palavra Vênus”mars exploration” +domain:edu localiza páginas sobre exploração de Marte provenientes, exclusivamente, de sites educacionais americanos
Buscar URL	url:u:allinurl:inurl:	Localiza páginas que possuam uma palavra ou frase específica na URL	url:jardim localiza páginas que contenham em sua URL a palavra jardim
Buscar <i>links</i>	link:	Localiza páginas que possuam links apontando para uma página ou domínio particular	link:www.unicamp.br localiza páginas com links apontando para o site da Unicamp
Buscar no título	title:allintitle:intitle:t:	Localiza páginas que possuam uma palavra ou frase específica no título	title:”comércio internacional” localiza páginas cujo título contenha a frase comércio internacional
Terminação	*	Localiza palavras com diferentes terminações	femini* localiza páginas que contenham as palavras feminina, femininas, feminino, feminilidade, etc.

Quadro 4 – Refinamento da busca
Fonte: Branski, 2004.

5.5 As livrarias eletrônicas

O avanço da *web* tem permitido que as empresas utilizem de novos serviços e os ofereçam aos usuários, na grande maioria com o objetivo de facilitar o contato destes com os seus produtos. Essa nova forma de comércio, definido por autores como comércio eletrônico, tem sido vista como uma nova ferramenta de divulgação para os seus produtos das empresas, e tem ganhado destaque na área.

O uso do comércio eletrônico vem beneficiando as empresas em vários aspectos, tais como: a agilidade das vendas, divulgação de seus produtos e o tempo de negociação reduzido. Essas facilidades encontradas pelos usuários fazem com que essa área se torne crescente e ganhe novos adeptos com o passar do tempo.

As livrarias encontraram na *web* essas novas formas de divulgar seus produtos, buscando atender maior número de pessoas. A livraria *Amazon* foi a primeira a estabelecer um endereço na *web*, sendo seguida por outras tantas livrarias. No Brasil, as livrarias de grande porte e nacionalmente conhecidas também buscaram esse novo instrumento para atrair os usuários, como por exemplo, a Livraria Saraiva e a Livraria Cultura.

Essas duas últimas livrarias citadas, as quais servirão de instrumentos de estudo deste trabalho, além de páginas disponíveis na *internet*, possuem suas lojas fisicamente nos grandes centros comerciais do país. Na grande maioria, as páginas na *web* apresentam diferentes formas de se realizar a pesquisa por algum produto, entre elas a busca livre e a busca controlada. A busca livre pode ser definida como aquela em que o usuário digita a informação desejada em um campo específico com os termos que lhe convém. Na busca controlada, a pesquisa é feita por meio da navegação pela taxonomia já definida pelo *site*.

5.5.1 Livraria Saraiva

A livraria Saraiva dispõe de 45 lojas em todo Brasil, sendo que 33 dessas lojas são consideradas *megastores*, ou seja, vendem diversos produtos, tais como: livros, Cds, DVDs, equipamentos de informática, materiais de papelaria entre outros.

A Saraiva ainda é uma das importantes editoras no país, abrangendo uma rede nacional de 12 filiais. Além desses serviços, a livraria Saraiva é considerada uma das pioneiras no mercado eletrônico, iniciando esse serviço a seus clientes no ano de 1998 (SARAIVA, *site*).

Na *web*, a Saraiva disponibiliza um *site*⁸ no qual seus clientes podem realizar compras dos produtos ou apenas realizar pesquisas de preços dos livros, por exemplo. Com o objetivo de facilitar o acesso dos usuários finais, o *site* disponibiliza seus produtos organizados por categorias e ainda, um campo de busca.

5.5.1.1 Estratégia de busca no *site* da livraria Saraiva

A busca por algum produto pode ser realizada tanto pela taxonomia construída para o *site*, que ilustra todas as seções estruturadas com o objetivo de agrupar os produtos que apresentam características semelhantes, quanto utilizando a busca livre, a qual compreende uma caixa de texto para o usuário digitar o produto que deseja, que se localiza na parte superior da tela, no item buscar em (figura 4).

⁸ Livraria Saraiva: www.saraiva.com.br



Figura 4 – Interface do site da livraria Saraiva.

A pesquisa livre realizada no site utilizando-se o campo “buscar em” permite que o próprio usuário utilize palavras-chaves de acordo com o assunto desejado. Porém, se desejar fazer uso de algumas ferramentas para refinar a busca, nem todas as operações são possíveis, como o uso de alguns operadores booleanos, as técnicas de proximidade e de truncamento.

A busca permite que o usuário utilize o operador booleano E, somente em português, não aceitando assim a variação AND deste operador. Esse fato pode ser comprovado conforme demonstra a figura 5, quando ao se realizar uma pesquisa por *livro E direito* temos um resultado e quando realizamos a busca por *livro AND direito* o site não nos traz documentos recuperados (figura 6).

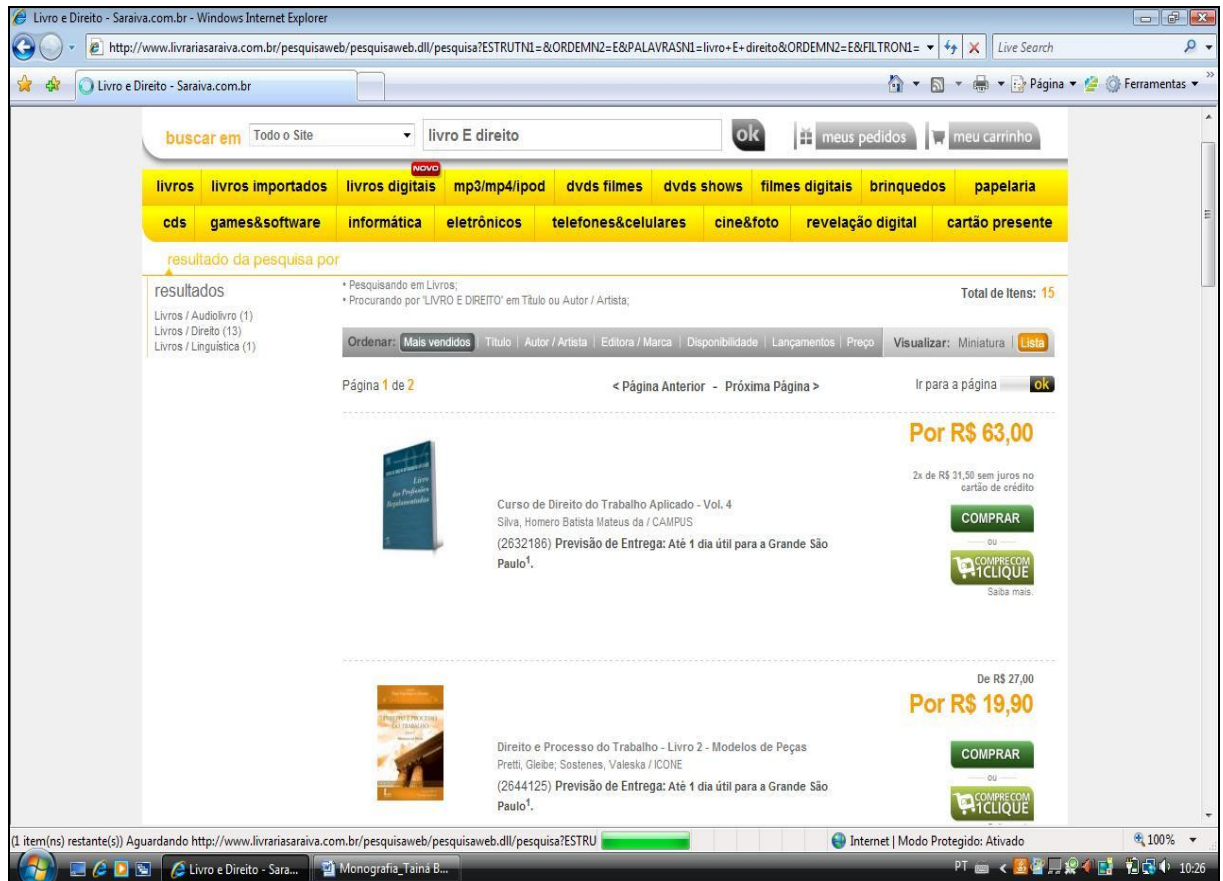


Figura 5 – Interface da pesquisa com operador booleano E no site da Saraiva.

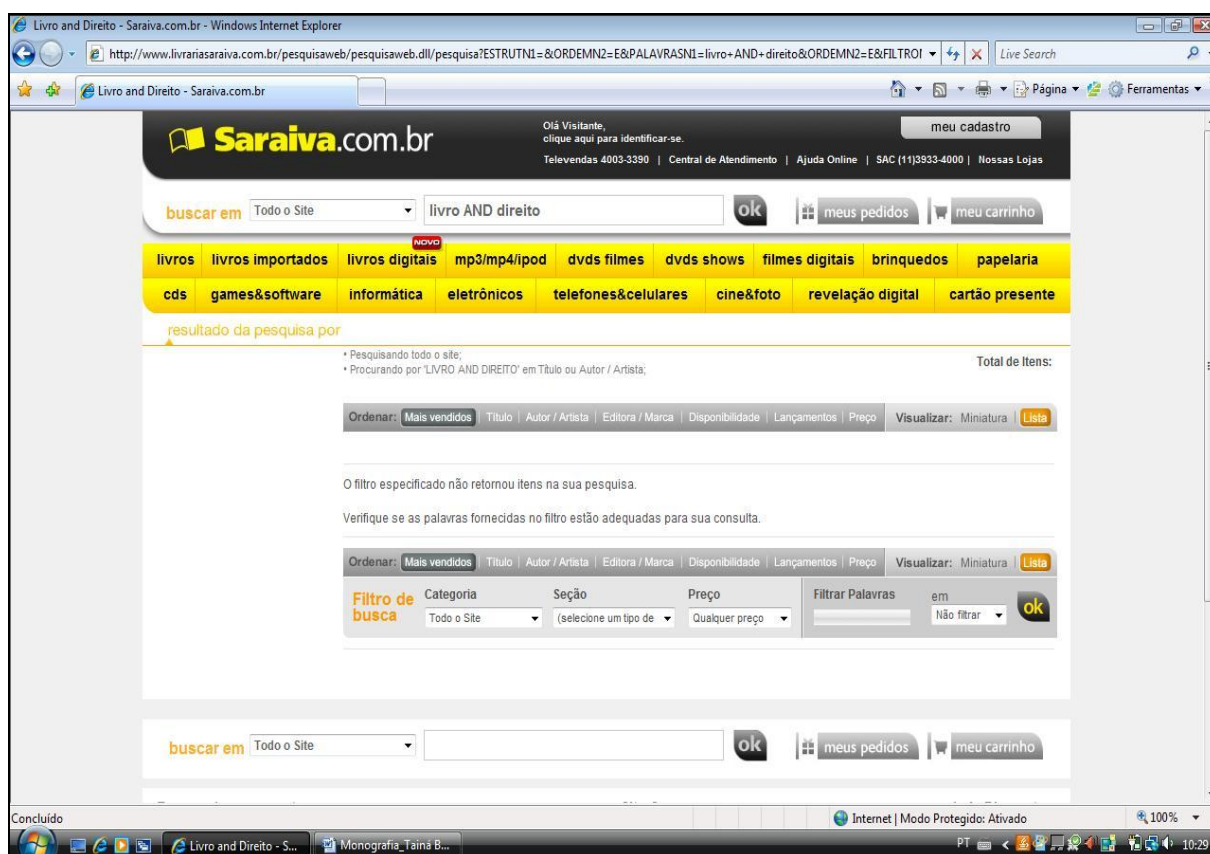


Figura 6 – Interface da pesquisa com operador booleano AND no site da Saraiva.

Porém, os resultados alcançados não podem ser generalizados para as próximas pesquisas, já que se utilizando do exemplo anterior e realizando novas buscas, mas com assuntos diferentes o resultado apresentado não foi satisfatório, como por exemplo, uma pesquisa com *livro E biblioteconomia*. Há livros que se referem a esses termos, mas não foram recuperados, como demonstra a figura 7. De acordo com a figura 8 é possível observar que não houve recuperação de documentos por meio da pesquisa pelo operador E.

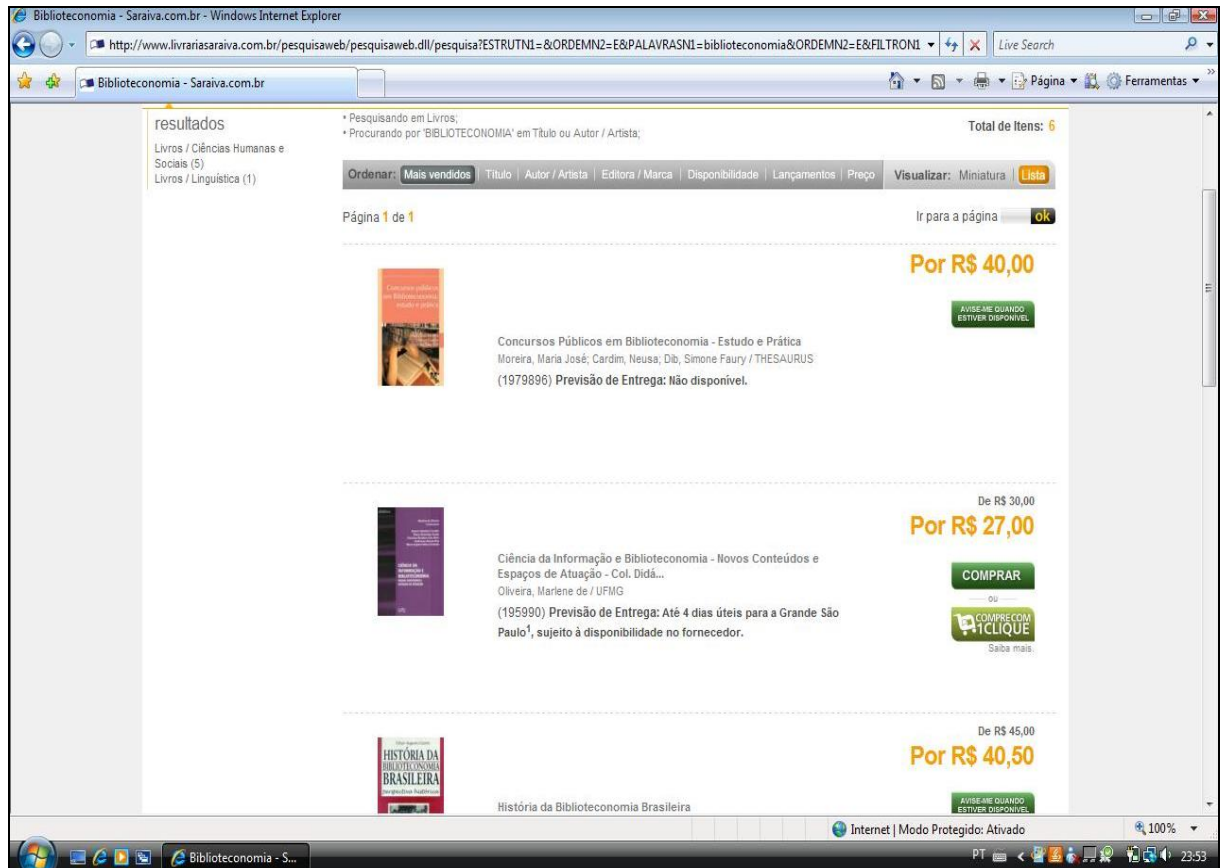


Figura 7 – Livros de Biblioteconomia na livraria Saraiva.

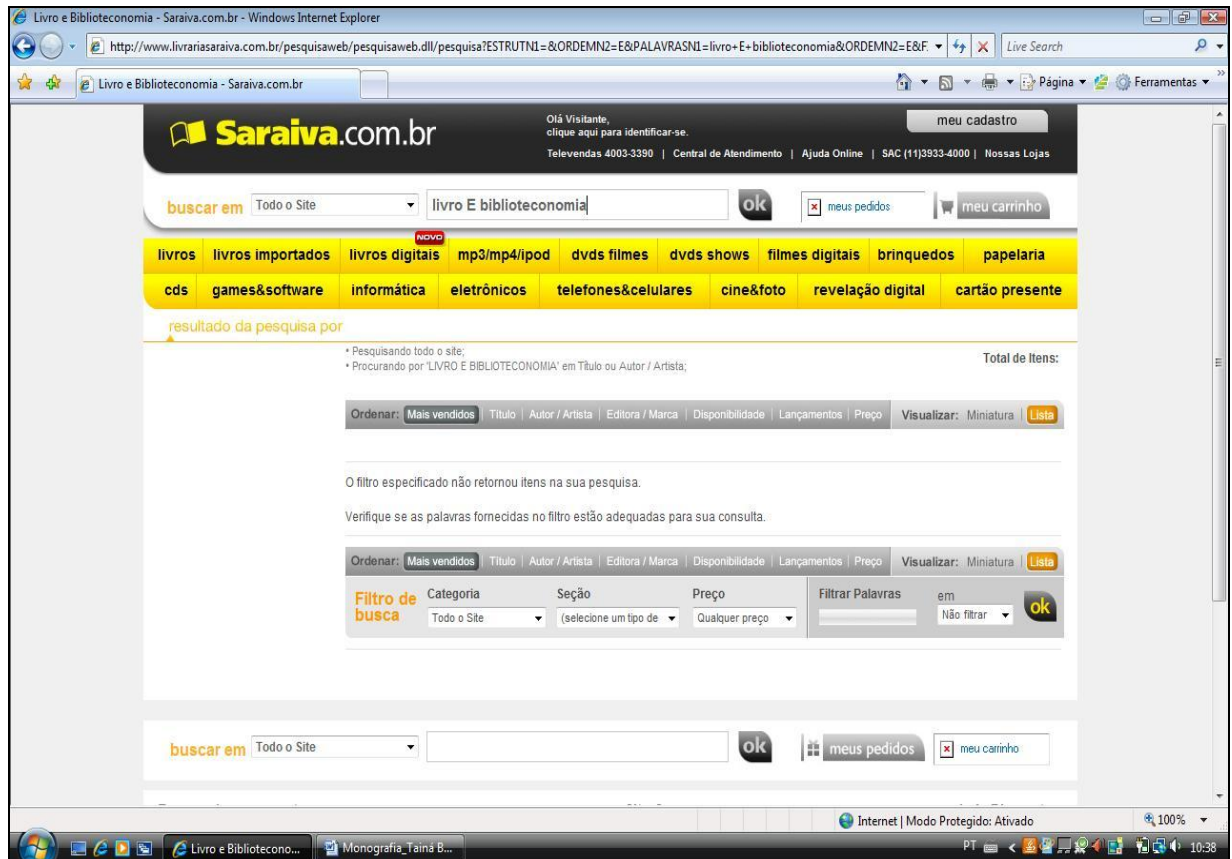


Figura 8 – Interface da pesquisa no site da Saraiva com resultado divergente da anterior.

A utilização de aspas na pesquisa demonstrou resultados satisfatórios.

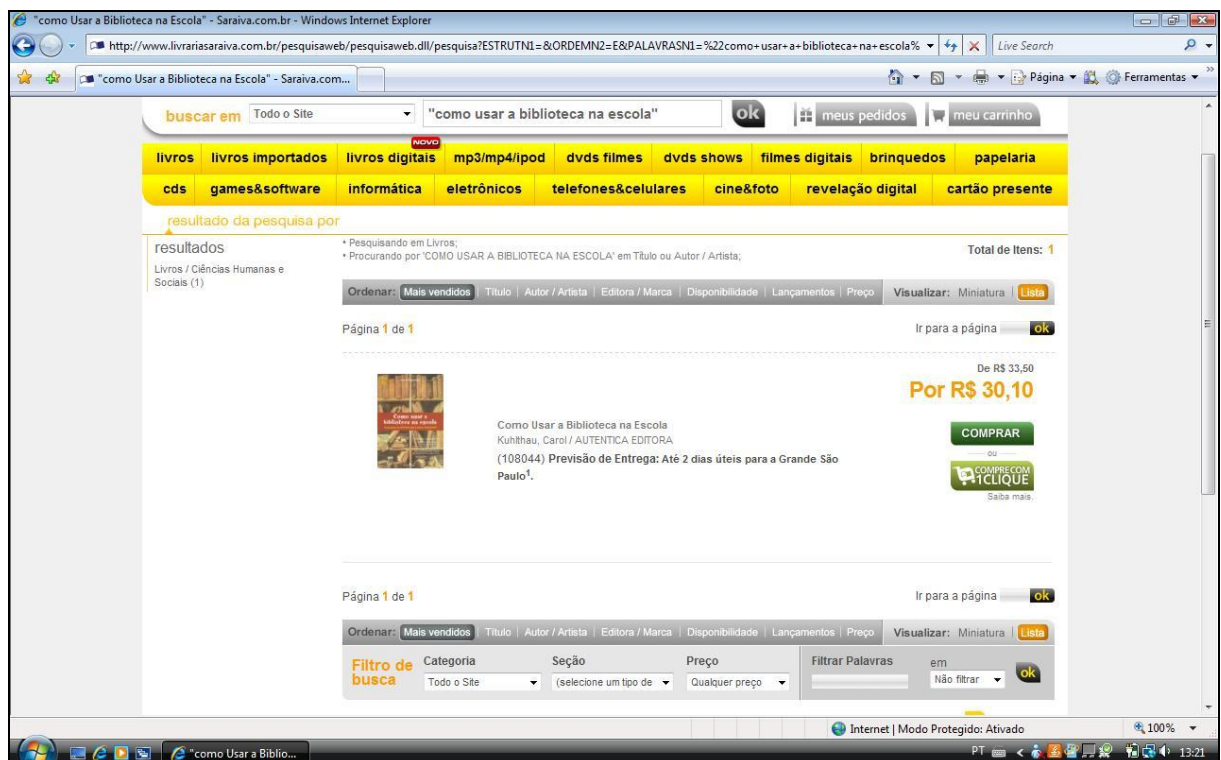


Figura 9 – Interface da pesquisa com a utilização das aspas no site da Saraiva.

Com o objetivo de verificar se os resultados obtidos com a busca livre ou a taxonomia apresentavam resultados satisfatórios com relação aos resultados, realizamos uma pesquisa comparativa utilizando o termo *biblioteconomia* para as possíveis comparações. No primeiro momento o termo foi pesquisado pela ferramenta de busca livre, recuperando ao total 6 itens, nos quais o termo encontrava-se no título de cada livro recuperado na busca (Figura 10). Já no segundo momento, com o uso da taxonomia foi possível recuperar 97 itens, porém nem todos os documentos estavam relacionados com o termo desejado. Muitos desses documentos estão relacionados com a área da arquivologia (Figuras 11 e 12).

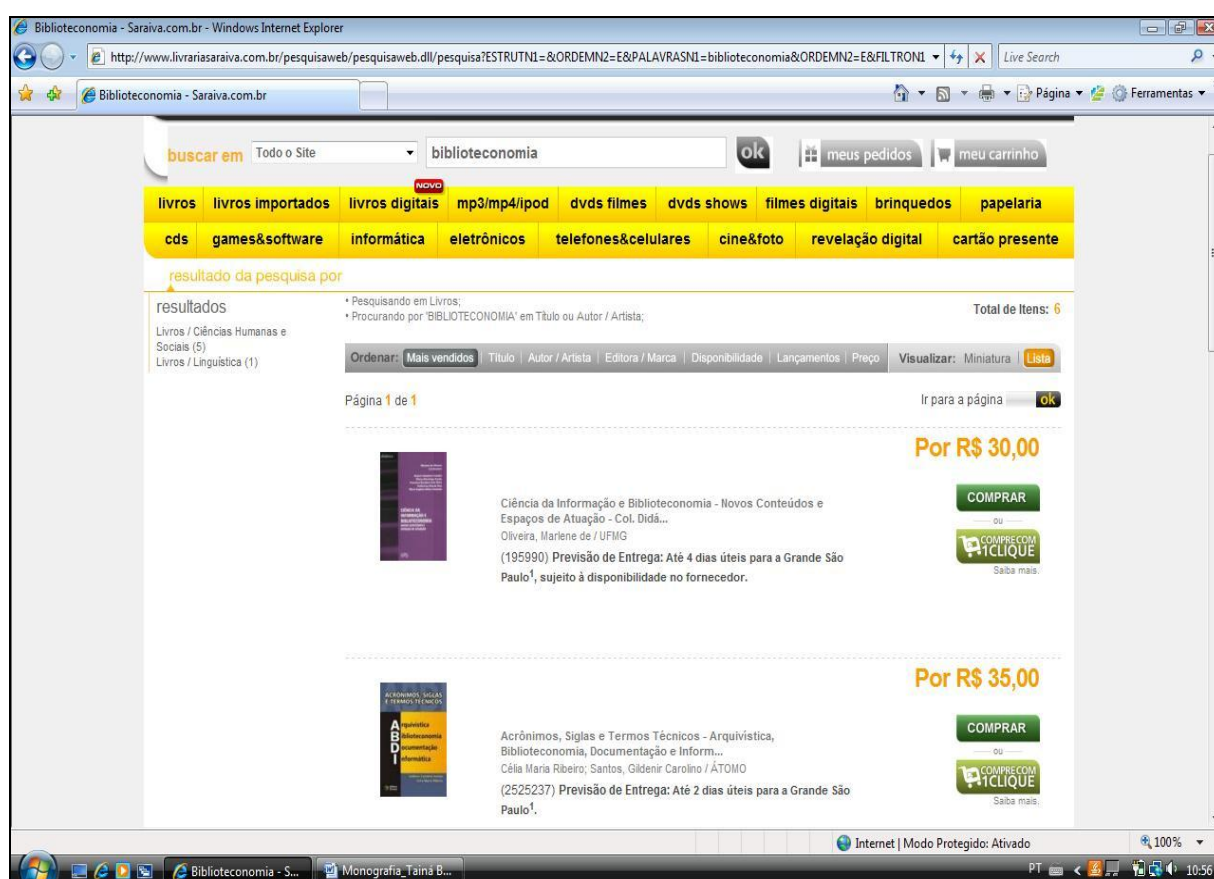


Figura 10 – Interface do resultado da busca por assunto no site da Saraiva.

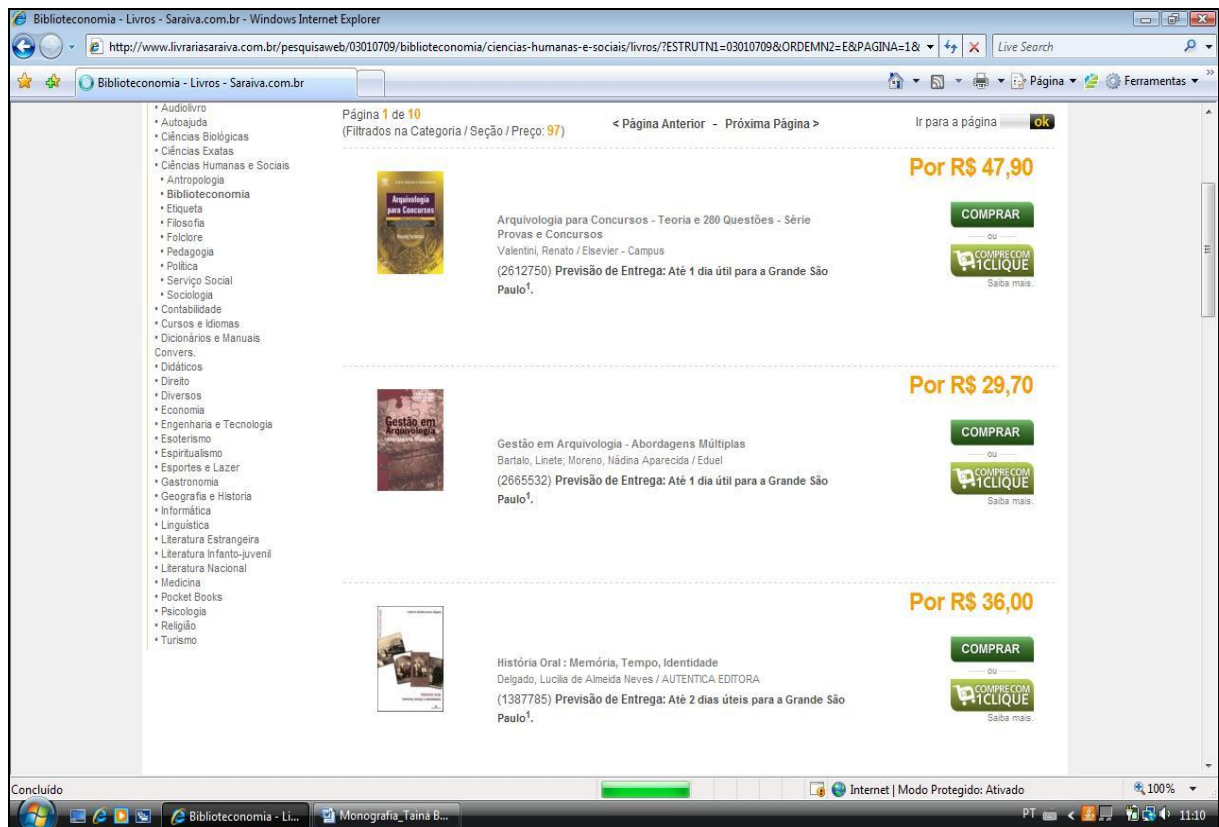


Figura 11 – Interface do resultado da busca no site livreria Saraiva que inclui os livros de Arquivologia.

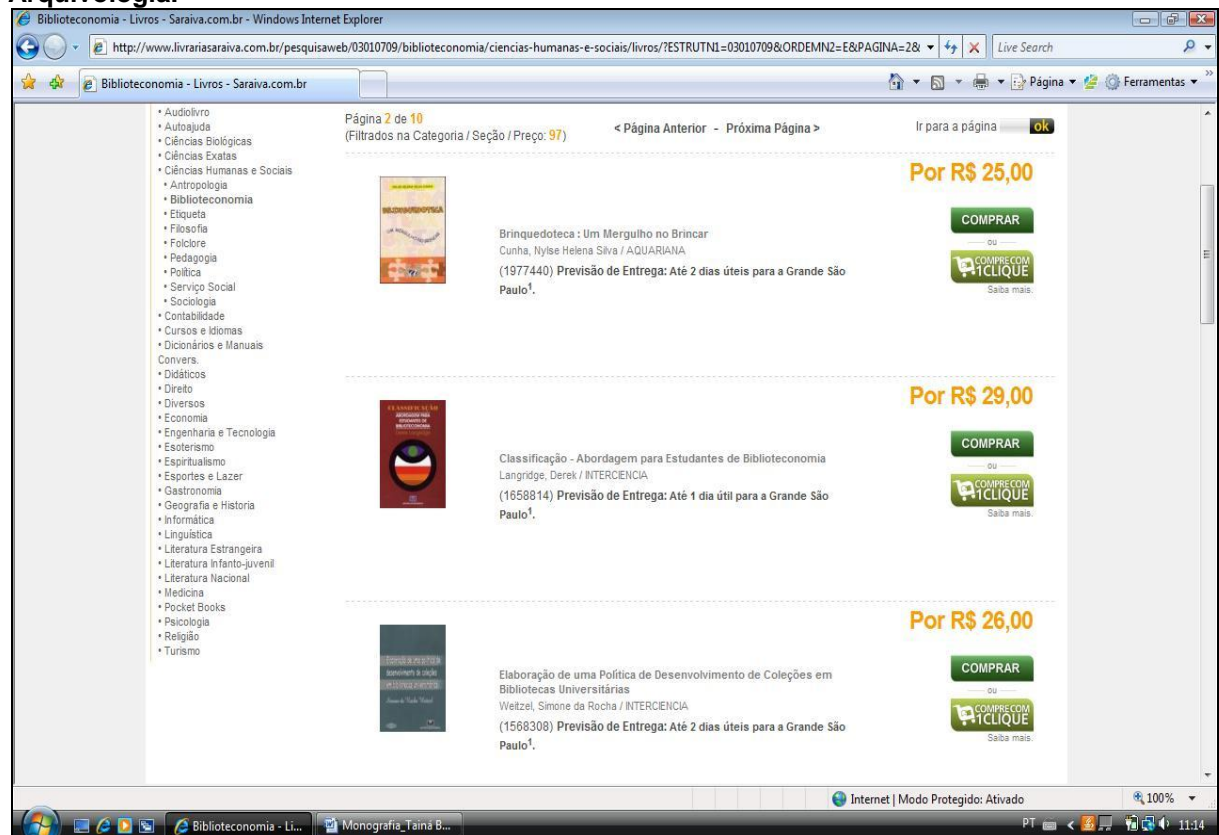


Figura 12 – Interface do resultado da busca no site da Saraiva com os livros de Biblioteconomia na segunda página.

5.5.2 Livraria Cultura

A livraria Cultura, fundada em 1947 por Eva Herz, mãe do atual presidente da instituição, é considerada uma referência para as demais empresas do mesmo setor. Há mais de 60 anos no mercado nacional, a livraria Cultura apresenta em seu catálogo uma grande variedade de produtos contendo mais de 3 milhões de títulos de livros. A visão da empresa representa de forma clara os seus objetivos

a Livraria Cultura acredita no poder transformador da informação. Queremos ser a melhor loja de informação e entretenimento e nos consolidar como a referência do setor. Quanto mais crescermos, mais distribuiremos informação e assim ajudaremos as pessoas a construir um mundo melhor e mais justo. (Livraria Cultura, 2010).

A Cultura também disponibiliza uma página na *web*⁹ a qual permite aos usuários e clientes da loja navegar pelos produtos. Na página é possível encontrar e comprar os diversos produtos que a empresa dispõe e ainda, informações gerais sobre as lojas.

5.5.2.1 Estratégia de busca no *site* da livraria Cultura

O usuário que desejar fazer uma pesquisa pelo *site* por algum produto encontrará duas maneiras semelhantes ao *site* da livraria Saraiva: utilizando a busca livre disponível pelo campo *busca* no alto da tela ou pesquisar navegando na taxonomia determinada pelo próprio *site*, no lado esquerdo da tela, como apresentado na figura 13.

⁹ Livraria Cultura: www.livrariacultura.com.br

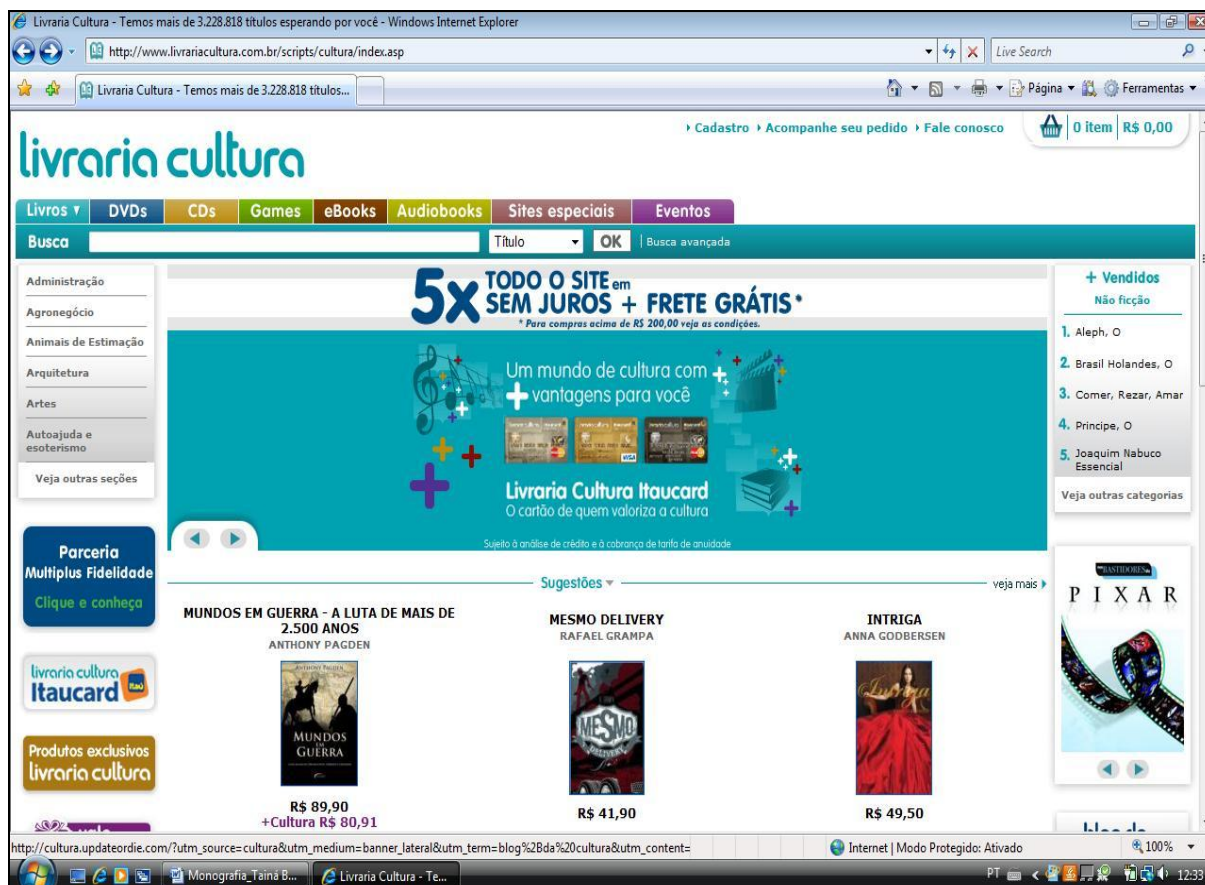


Figura 13 – Interface do site da livraria Cultura.

A busca controlada permite que o usuário navegue pelas seções estabelecidas pelo site até que a informação seja encontrada. A maior dificuldade que um usuário comum pode encontrar é em identificar em qual categoria o assunto que deseja está inserido. Esse fato pode tornar a pesquisa demorada.

A pesquisa livre no site encontra-se disponível no campo *busca*, e pode ser refinada pelo título, autor, editora, ISBN, coleção ou *tag*. Entretanto, o usuário que deseja utilizar os operadores booleanos ou outra estratégia de busca encontrará alguns obstáculos, pois o site não permite ao usuário fazer uso desses instrumentos. Como exemplo, ao realizar uma pesquisa utilizando os termos *biblioteconomia AND livro* (Figura 14) não há recuperação de documentos.

Livraria Cultura - Temos mais de 3.229.550 de títulos esperando por você - Windows Internet Explorer

http://www.livriacultura.com.br/scripts/cultura/busca/busca_avancada.asp?tipoBusca=2

Livraria Cultura - Temos mais de 3.229.550 de títulos...

Cadastro Acompanhe seu pedido Fale conosco 0 item R\$ 0,00

livraria cultura

Livros DVDs CDs Games eBooks Audiobooks Sites especiais Eventos

Busca biblioteconomia AND livro Título OK Busca avançada

A informação solicitada foi pesquisada, mas não foi encontrado o livro correspondente.

Clique [aqui](#) e veja nossas dicas para realizar a busca avançada.

Nosso catálogo contém hoje 2.888.964 livros, entre títulos nacionais e importados. Você pode preencher um ou mais campos para realizar a busca.

BUSCA AVANÇADA

Título:

☒ Palavra(s) do título ☐ Início do título

Autor:

☒ Palavra(s) do nome ☐ Início do sobrenome

Editora:

Coleção:

ISBN:

Refine sua busca:

Assunto: Qualquer assunto

Idioma: Qualquer idioma

Faixa de preço: Qualquer preço

Internet | Modo Protegido: Ativado 100%

PT 09:24

Figura 14 – Interface da pesquisa com operador booleano AND no site da Cultura

O mesmo acontece ao se pesquisar pelo mesmo assunto anterior, com o operador booleano em português, *biblioteconomia E livro*, como por exemplo (figura 15). Da mesma forma, o usuário é remetido a uma página de pesquisa avançada.

The screenshot shows the Livraria Cultura website in a Windows Internet Explorer browser. The address bar displays the URL: http://www.livriacultura.com.br/scripts/cultura/busca/busca_avancada.asp?tipoBusca=2. The website header includes the logo "livraria cultura" and navigation links: "Cadastro", "Acompanhe seu pedido", "Fale conosco", and a shopping cart icon showing "0 item R\$ 0,00". A horizontal menu bar contains categories: "Livros", "DVDs", "CDs", "Games", "eBooks", "Audiobooks", "Sites especiais", and "Eventos". Below this is a search bar with the text "Busca biblioteca E livro" and a dropdown menu set to "Título". A green "OK" button and a link to "Busca avançada" are also present. The main content area displays a message: "A informação solicitada foi pesquisada, mas não foi encontrado o livro correspondente." followed by a link "Clique aqui e veja nossas dicas para realizar a busca avançada." Below this, it states: "Nosso catálogo contém hoje 2.888.964 livros, entre títulos nacionais e importados. Você pode preencher um ou mais campos para realizar a busca." The "BUSCA AVANÇADA" section contains several input fields: "Título:", "Autor:", "Editora:", "Coleção:", "ISBN:", "Assunto:" (with a dropdown menu set to "Qualquer assunto"), "Idioma:" (with a dropdown menu set to "Qualquer idioma"), and "Faixa de preço:" (with a dropdown menu set to "Qualquer preço"). The browser's status bar at the bottom shows "Internet | Modo Protegido: Ativado", "100%", and the time "20:16".

Figura 15 – Interface da pesquisa com operador booleano E no site da Cultura

Esse resultado da pesquisa também pode ser obtido ao se tentar pesquisar utilizando qualquer outra técnica para refinar a busca. Ao se utilizar as aspas na pesquisa, o resultado também se apresentou satisfatório na recuperação do produto, semelhante na livraria Saraiva, como demonstra a figura 16.

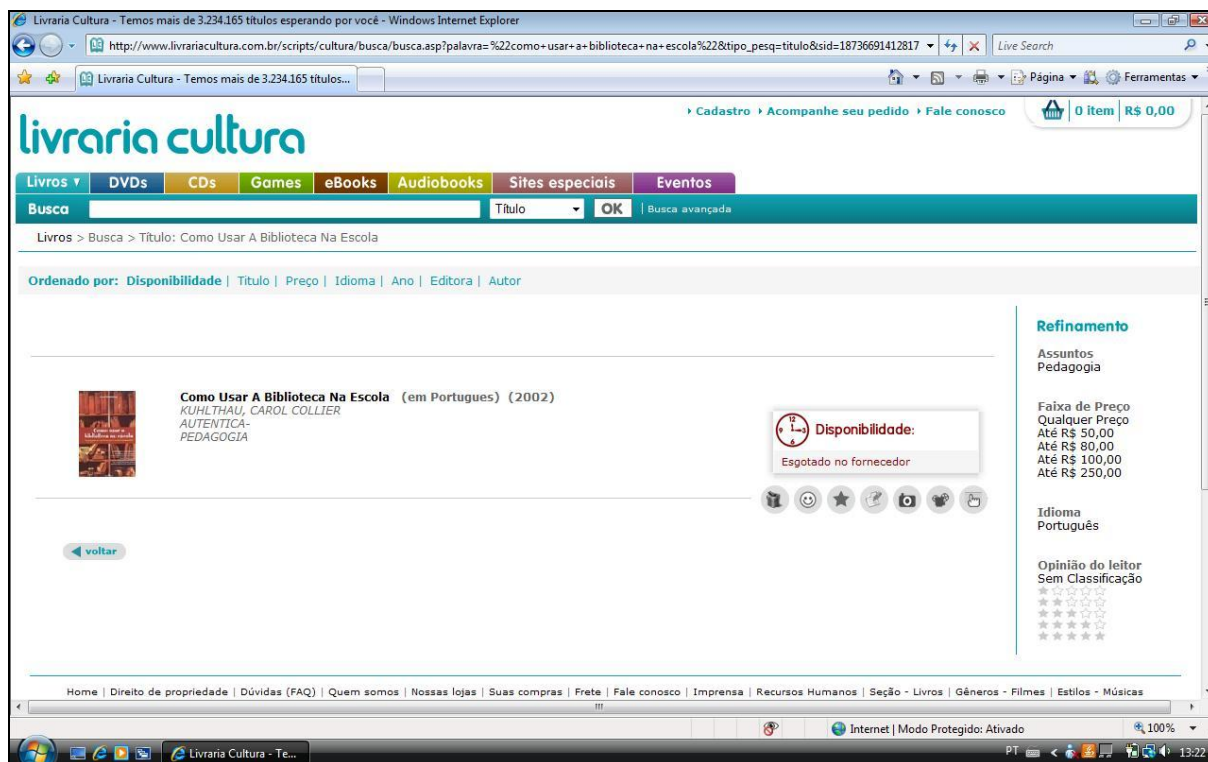


Figura 16 – Interface do resultado da busca com o uso de aspas.

6 METODOLOGIA

6.1 Introdução

Apesar do grande avanço dos recursos disponíveis na *internet* e das ferramentas que auxiliam os sistemas de recuperação da informação algumas divergências nos resultados ainda são encontradas quando se opta a utilizar a linguagem natural ou a controlada. Com isso, o motivo desta pesquisa é analisar e fazer as devidas comparações dos resultados obtidos fazendo uso da linguagem natural ou das taxonomias usadas nos próprios *sites* das livrarias eletrônicas pelos usuários.

As livrarias eletrônicas utilizadas para a aplicação desse estudo foram: a Livraria Saraiva e a Livraria Cultura, já descritas na revisão de literatura deste trabalho. A escolha se deve ao fato de cada uma dessas compreenderem uma rede significativa no comércio nacional.

Dentre os critérios apresentados na revisão de literatura deste trabalho, para se formar um quadro de características tanto da linguagem natural quanto da linguagem controlada, escolhemos dois desses para direcionar a pesquisa e que possivelmente a análise destes responderá o problema deste estudo. Os critérios escolhidos foram quanto ao tempo e os resultados obtidos na pesquisa.

Nesta etapa da pesquisa serão apresentados ainda, a escolha da metodologia, os instrumentos utilizados para a coleta de dados e a população desta pesquisa.

6.2 A metodologia

A segunda etapa deste trabalho, sendo a primeira formada pela revisão bibliográfica, consiste em uma pesquisa exploratória de natureza quantitativa, na qual foi aplicado ao usuário um questionário que abordava questões referentes à busca nos *sites* das livrarias eletrônicas.

Sobre a pesquisa quantitativa, Baptista e Cunha (2007) esclarecem

a pesquisa quantitativa caracteriza-se, tanto na fase de coleta de dados quanto no seu tratamento, pela utilização de técnicas estatísticas. (...) O seu uso intensivo teve por objetivo garantir uma maior precisão na análise e interpretação dos resultados, tentando, assim, aumentar a margem de confiabilidade quanto às inferências dos resultados encontrados. (BAPTISTA; CUNHA, 2007, p. 173).

Nesse tipo de pesquisa, os dados coletados são estruturados estatisticamente para serem apresentados. Hübner (1998, p. 56) explica que “o que define uma pesquisa como sendo qualitativa ou quantitativa não é o método de coleta, mas sim a forma de tratamento dos dados”.

A escolha da metodologia deve estar de acordo com a proposta da pesquisa, buscando obter dados para serem analisados e posteriormente responder a pergunta problema deste estudo.

6.3 Coleta de dados e a população da pesquisa

Os dados obtidos para esta pesquisa foram coletados diretamente e pela primeira vez pelo responsável, ou seja, são considerados dados primários da pesquisa. Para a coleta, foi utilizado previamente um roteiro (Apêndice I), o qual auxiliava o entrevistado para o processo de recuperação da informação e, após esse primeiro passo, foi aplicado um questionário (Apêndice II) a fim de avaliar a pesquisa realizada nos *sítes* das livrarias eletrônicas.

Na maioria das entrevistas, o entrevistador esteve presente até a conclusão da pesquisa e preenchimento do questionário pelo sujeito participante. Nos outros casos, os questionários foram encaminhados via correio eletrônico, porém antes do envio foi explicado aos entrevistados como deveriam responder o instrumento de pesquisa.

Segundo Baptista e Cunha (2007) o questionário “é um dos métodos mais utilizados. Consiste numa lista de questões formuladas pelo pesquisador a serem respondidas pelos sujeitos pesquisados”.

A coleta de dados foi realizada com 30 pessoas comuns, usuários da *internet* ou não, escolhidos de forma aleatória. A população deste estudo é constituída por usuários de diferentes faixas etárias e com diferentes ocupações.

A busca pela informação nos *sítes* das livrarias eletrônicas, como apresentado na revisão de literatura deste trabalho, pode ser feita por duas maneiras distintas: utilizando-se os termos dispostos em uma lista de assunto categorizada e hierarquizada pelo próprio *site* ou utilizar o campo em que é possível digitar diretamente o que se deseja encontrar (*buscar em* – Livraria Saraiva ou *busca* – Livraria Cultura).

Para a pesquisa, foi definido um produto igual em todos os dois *sítes* pelo qual o entrevistado deveria realizar a busca. O produto escolhido, aleatoriamente, foi um livro com as seguintes informações: título: Como usar a biblioteca na escola, autor: Carol Kuhlthau, ano: 2002, editora: Autentica-Pedagogia e assunto principal: Biblioteconomia.

Com o objetivo de haver melhor análise do comportamento de busca do usuário, foi decidido que não seria permitido o entrevistado buscar utilizando o título

principal do livro, mas sim utilizando palavras-chaves, outras técnicas de pesquisa ou ainda, as informações do livro contidas no roteiro.

6.4 Questionário

O questionário utilizado como instrumento para essa pesquisa é composto por questões fechadas, divididas em itens e subitens a serem analisados pelo entrevistador. Apenas o último subitem referente a cada livraria se caracteriza por ser uma questão aberta optativa ao entrevistado.

O instrumento apresentado se aplica aos dois *sites* de livrarias eletrônicas, com apenas pequenas alterações em alguns termos, tais como *buscar em* ou *busca*, para que se tornasse coerente com a página a ser avaliada.

6.5 Atividade proposta ao entrevistado

Foi proposto ao entrevistado que realizasse a pesquisa no *site* da livraria eletrônica pelo produto determinado neste estudo, sem que fosse dito ao sujeito que havia duas formas de pesquisa, a busca livre e a busca controlada. Foi observado qual a primeira estratégia de busca utilizada pelo usuário, a qual deveria ser respondida e avaliada posteriormente no questionário.

No segundo momento da pesquisa, foi proposto que o usuário utilizasse o outro processo de busca no *site*, este exposto pelo entrevistador, mas sem maiores informações para que não houvesse influência no modo de pesquisa. Ao final desta etapa, o entrevistado também deveria responder às questões referentes no questionário e avaliar o processo.

Para que não ocorresse qualquer tipo de influência no modo de pesquisa nos dois *sites*, foi dividido anteriormente à aplicação do estudo prático, questionários que iniciassem a pesquisa pela livraria Saraiva e livraria Cultura, com o objetivo de haver maior precisão nos resultados.

7 ANÁLISE DOS DADOS

A análise dos dados coletados por meio do questionário aplicado aos entrevistados irá permitir que as conclusões posteriormente apresentadas estejam fundamentadas na experiência de cada sujeito da pesquisa.

As respostas estão baseadas nos 30 questionários respondidos pelos entrevistados. Dentre esses questionários, alguns apresentaram respostas para a questão aberta e que também serão apresentadas neste trabalho, com o objetivo de fazer uma avaliação a partir da opinião direta do sujeito entrevistado.

7.1 Perfil do entrevistado

A partir dos dados coletados dos 30 questionários aplicados, foi possível observar que o perfil dos entrevistados é de maioria composta por mulheres, correspondendo a 60% dos participantes, com predominância de idade entre a faixa etária de 15 a 25 anos, enquanto que os entrevistados do sexo masculino correspondem a 40% também com maioria na mesma faixa etária.

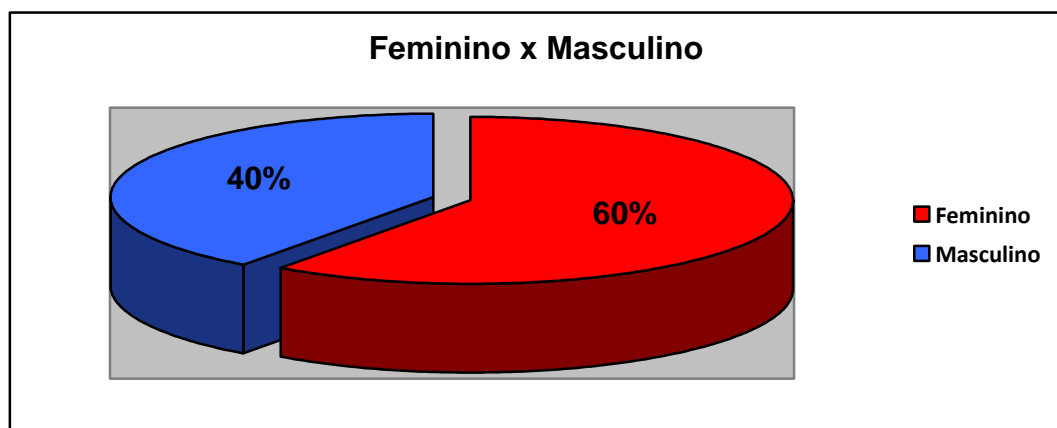


Gráfico 1 – Porcentual de entrevistados de acordo com o sexo

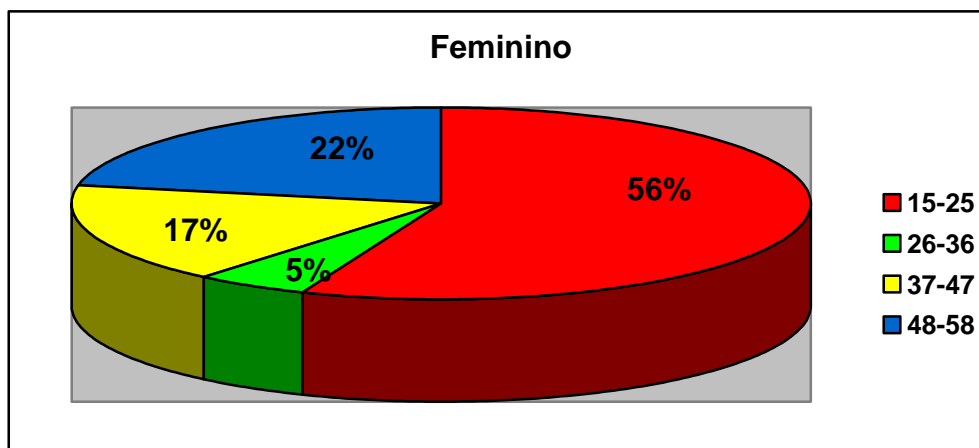


Gráfico 2 – Faixa etária entre entrevistados do sexo feminino

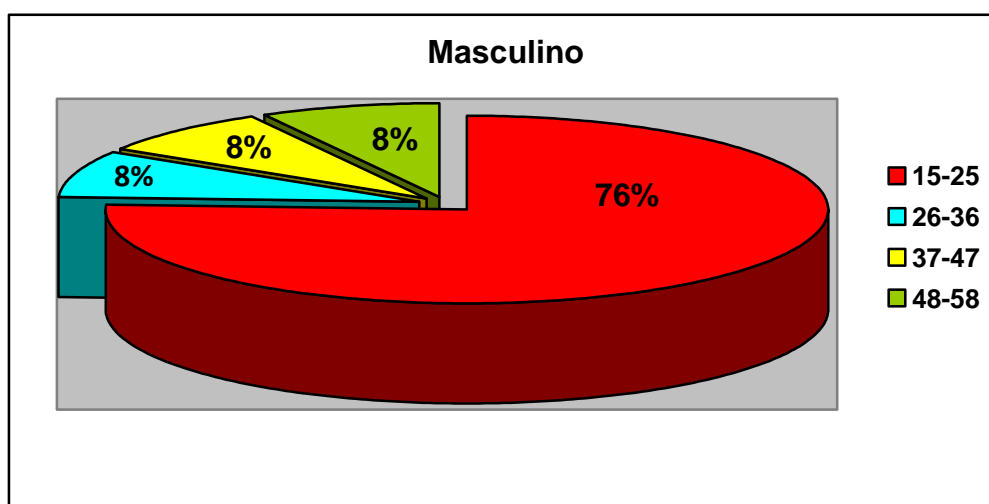


Gráfico 3 – Faixa etária entre entrevistados do sexo masculino

A maioria dos entrevistados é do sexo feminino, como visto anteriormente, e tem como profissão predominante serem estudantes universitárias, correspondendo a 45%. Os homens, na sua maioria, são também estudantes universitários, com o percentual de 34%.

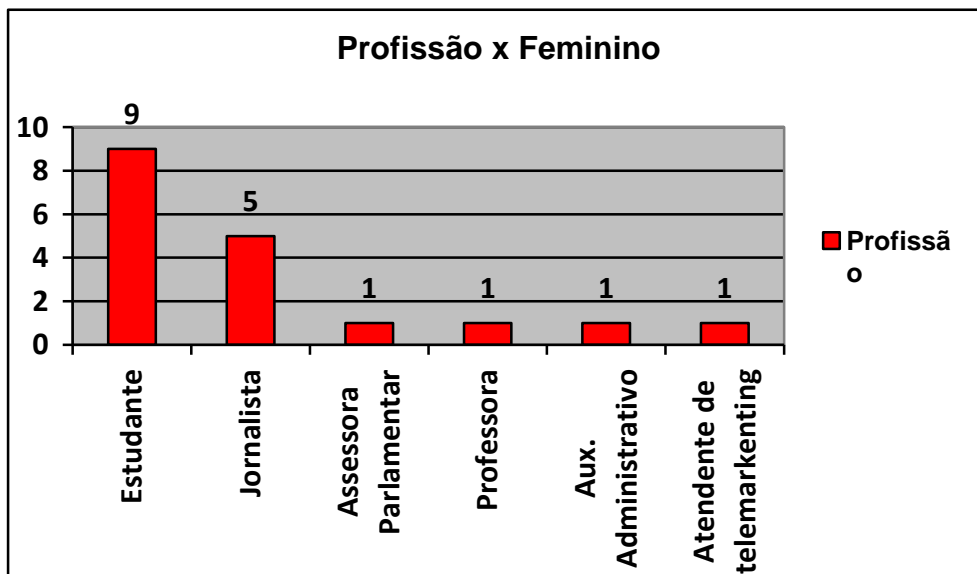


Gráfico 4 – Quantidade de entrevistados do sexo feminino quanto à profissão

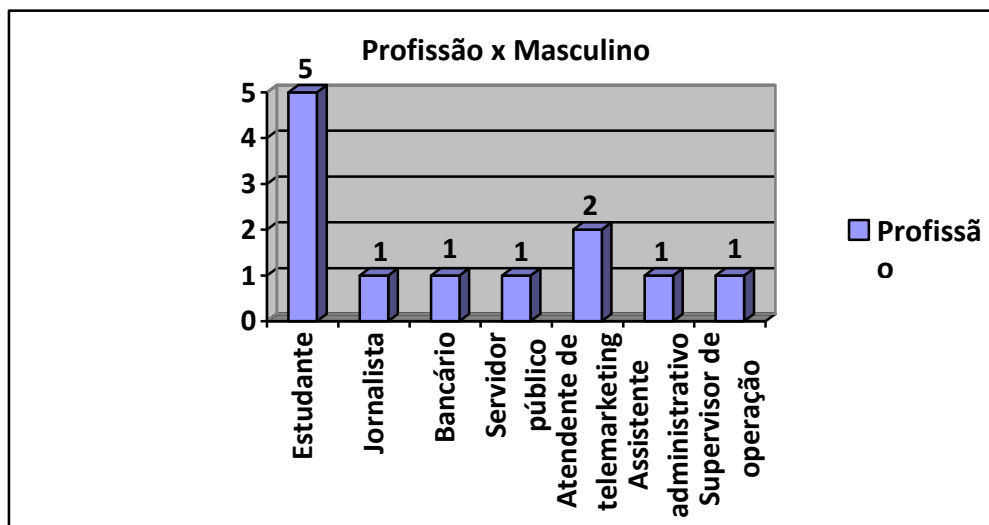


Gráfico 5 – Quantidade de entrevistados do sexo masculino quanto à profissão

Com relação à frequência que o sujeito faz uso da internet e também com relação ao uso desta para a realização de pesquisas, foi possível observar que do total de entrevistados 90% (gráfico 6) faz uso frequente e que, 77% fazem pesquisas na internet frequentemente (gráfico 7).

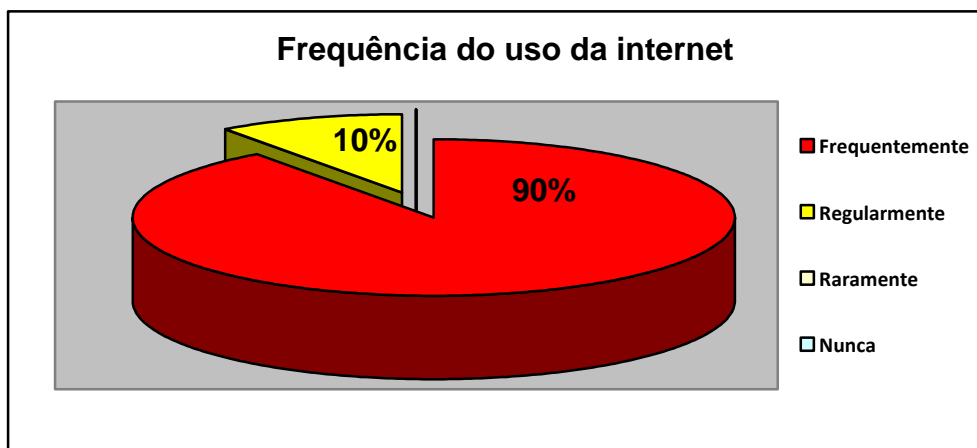


Gráfico 6 – Porcentagem dos entrevistados com relação à frequência do uso da internet

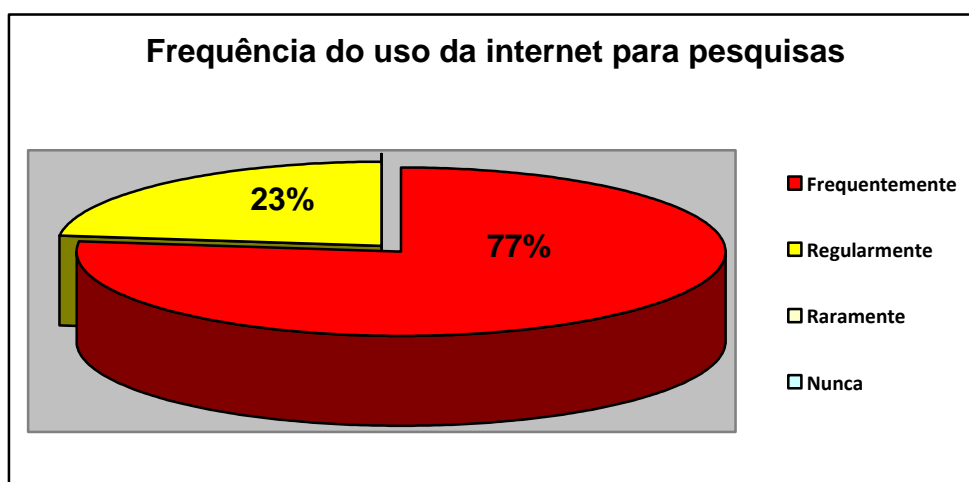


Gráfico 7 – Porcentagem dos entrevistados com relação à frequência de pesquisas na internet

7.2 Conhecimento sobre estratégias de busca

Nessa questão sobre o conhecimento de estratégias de busca para facilitar no processo de pesquisa, o entrevistado poderia marcar mais de uma opção de acordo com o seu conhecimento das técnicas. A partir das informações dos questionários, foi possível observar que 26% dos entrevistados conheciam e faziam uso de operadores booleanos, quanto que a maioria utiliza nas pesquisas as aspas, 41%. Aqueles que responderam que não conheciam e nem utilizavam qualquer tipo de instrumento para a realização de pesquisas, somam 21% dos entrevistados.

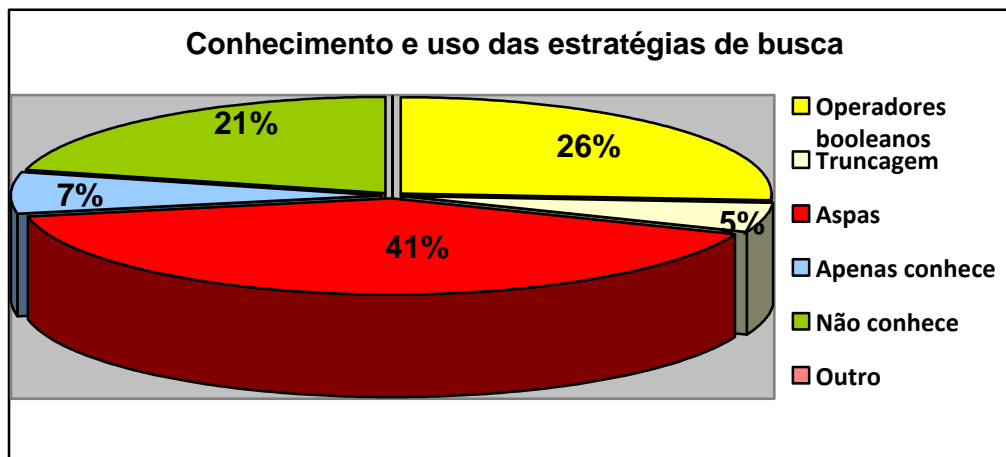


Gráfico 8 – Porcentagem dos entrevistados em relação ao conhecimento sobre estratégias de busca

De acordo com as respostas, ainda, foi possível observar que os mais jovens são os que mais utilizam estratégia de busca. O gráfico a seguir mostra a relação da quantidade de entrevistados por faixa etária que fazem uso de pelo menos uma estratégia de busca no momento da pesquisa.

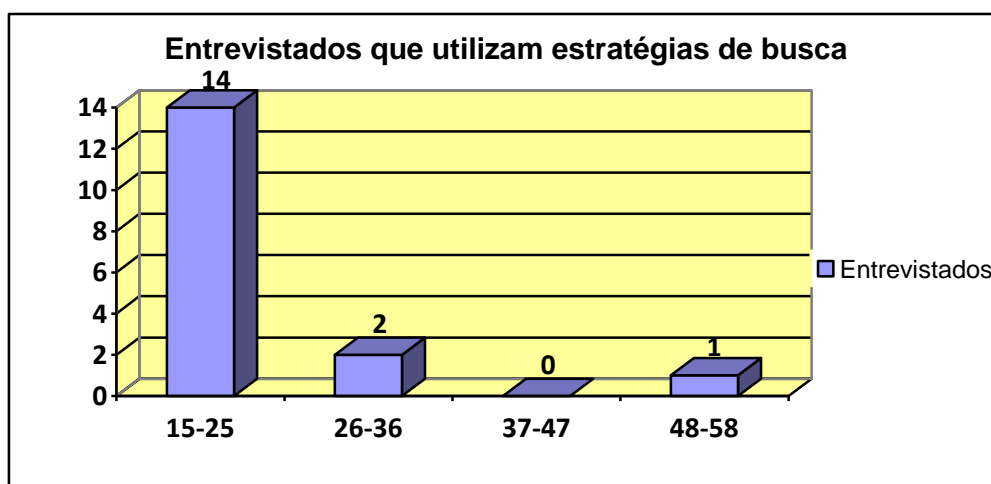


Gráfico 9 – Quantidade de entrevistados de acordo com a faixa etária que utilizam pelo menos uma estratégia de busca

7.3 Conhecimento dos sites da Livraria Saraiva e Cultura

Sobre as livrarias Saraiva e Cultura, escolhidas para a realização deste estudo, foi perguntado ao entrevistado se alguma vez já havia realizado pesquisa por algum produto nesses sites. Com os dados obtidos foi possível perceber que a

grande maioria diz ter realizado alguma busca nos dois *sites*, correspondendo a 67% do total dos sujeitos da pesquisa.

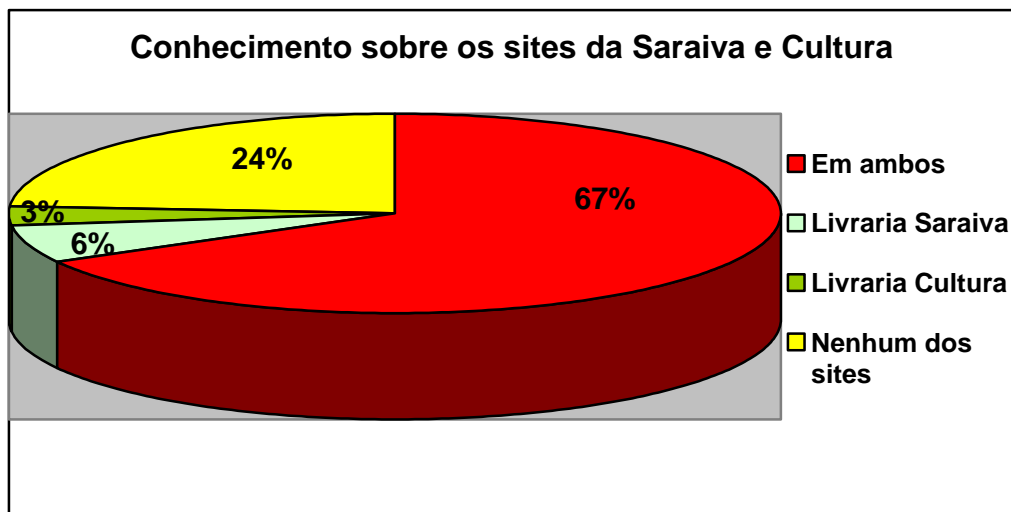


Gráfico 10 – Porcentagem dos entrevistados com relação o conhecimento dos *sites* das livrarias.

7.4 Sobre a Livraria Saraiva

Os gráficos nessa etapa da análise correspondem especificamente à livraria Saraiva. Com relação à primeira pesquisa, o caminho mais acessado pelos usuários foi o campo “buscar em”, no qual era possível inserir dados para realizar a pesquisa.

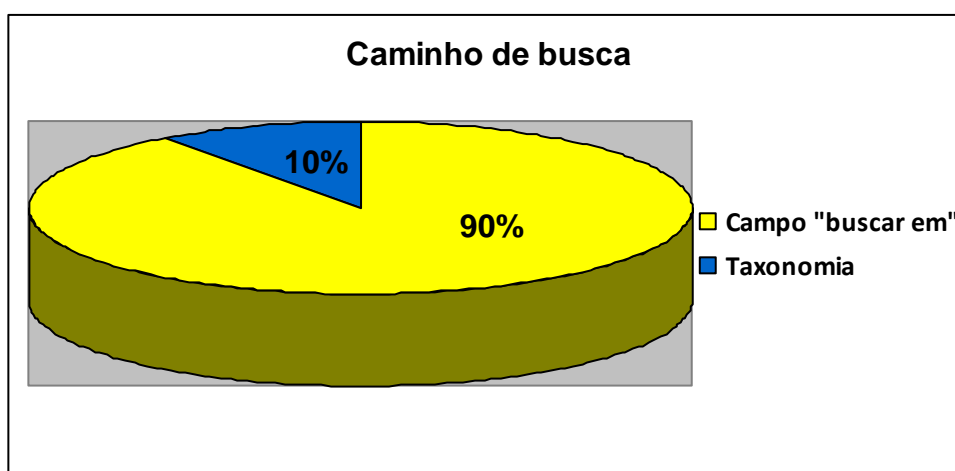


Gráfico 11 – Porcentagem do total de entrevistados com relação ao caminho utilizado para a busca

Ao fim da primeira pesquisa pelo produto no *site*, foi perguntado ao usuário sobre o resultado da pesquisa, se este foi satisfatório ou não. Dos que utilizaram o caminho “buscar em”, 22 entrevistados responderam que o resultado foi satisfatório. Os entrevistados que optaram por utilizar a taxonomia na primeira busca, não obtiveram sucesso na pesquisa, demonstrando dificuldade para tanto.

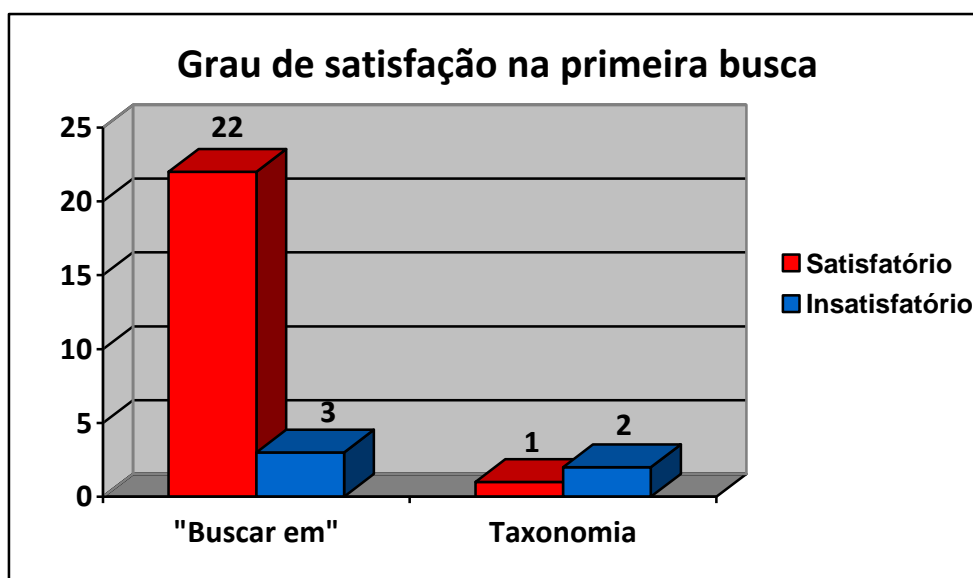


Gráfico 12 – Grau de satisfação quanto ao resultado da primeira pesquisa

Em relação às possíveis dificuldades encontradas durante a primeira pesquisa, 83% dos entrevistados não encontrou obstáculos para a realização da busca (gráfico 13). Deste porcentual, apenas 1 entrevistado utilizou a taxonomia para a pesquisa, enquanto o restante dos participantes utilizaram a busca livre.

Dos que tiveram dificuldades tem-se 5 entrevistados, o que corresponde a 17%. As dificuldades encontradas estão relacionadas à falta de clareza na taxonomia, encontrar o campo para iniciar a busca e outros.

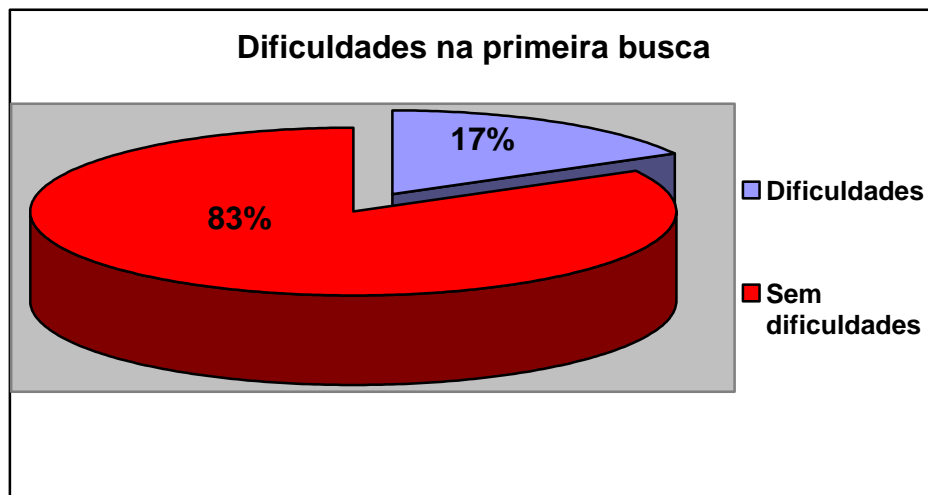


Gráfico 13 – Porcentagem dos entrevistados quanto às dificuldades



Gráfico 14 – Grau de dificuldades em relação aos itens expostos

Com relação às tentativas para se encontrar o produto no *site*, foi possível observar que a maioria daqueles que utilizaram o campo “buscar em” encontraram o livro em apenas 1 tentativa enquanto que os que utilizaram a taxonomia percorreram, na maioria 3 caminhos. Esses caminhos correspondem aos níveis das taxonomias, ou seja, se os entrevistados navegaram apenas no primeiro nível da taxonomia ou ainda utilizaram o segundo e mais níveis para a realização da pesquisa.

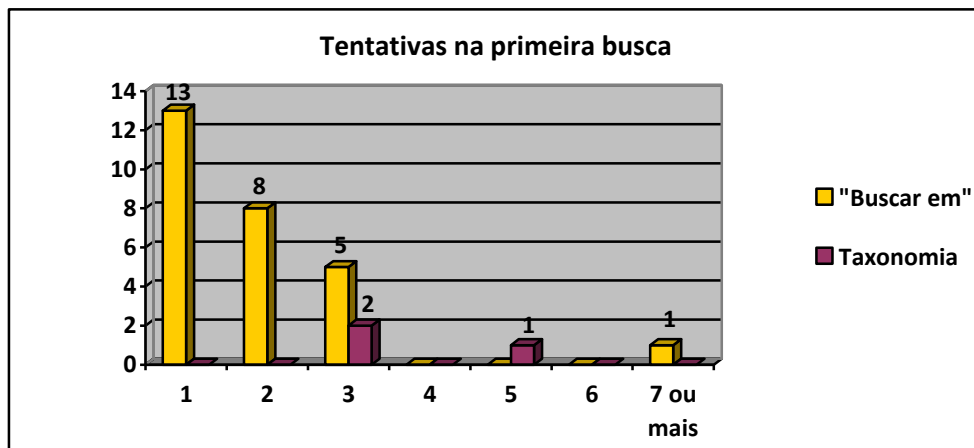


Gráfico 15 – Quantidade de tentativas para a primeira pesquisa

Depois de realizada a segunda pesquisa no *site* utilizando o outro modo de busca sem ser aquele já utilizado, foi questionado ao entrevistado, com o objetivo de comparar os dois modos de busca, se os resultados agora apresentados eram iguais a primeira pesquisa. A maioria, 77%, respondeu que o resultado foi diferente da primeira. Após os dois modos de busca propostos, os entrevistados responderam que o tempo de pesquisa e a quantidade de documentos recuperados são apontados como maior fator diferenciador dos dois modos.

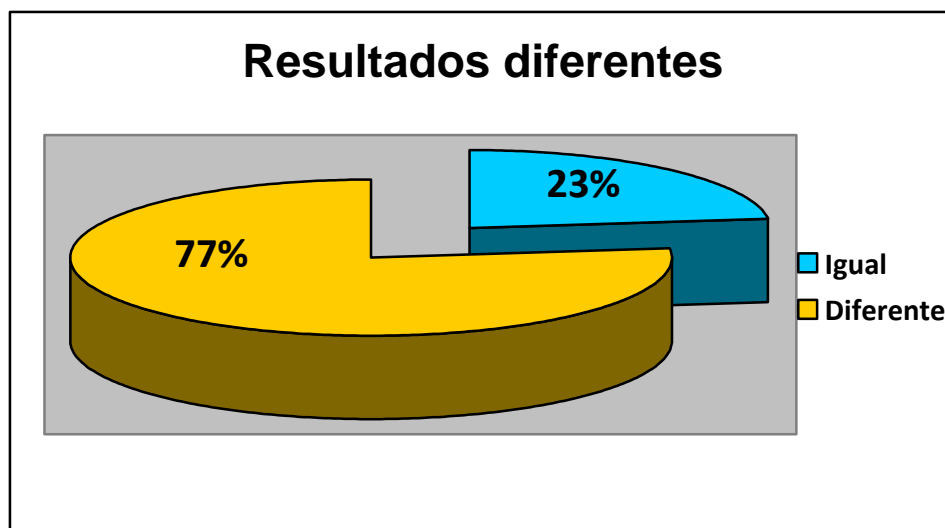


Gráfico 16 – Grau de diferenças entre os dois modos

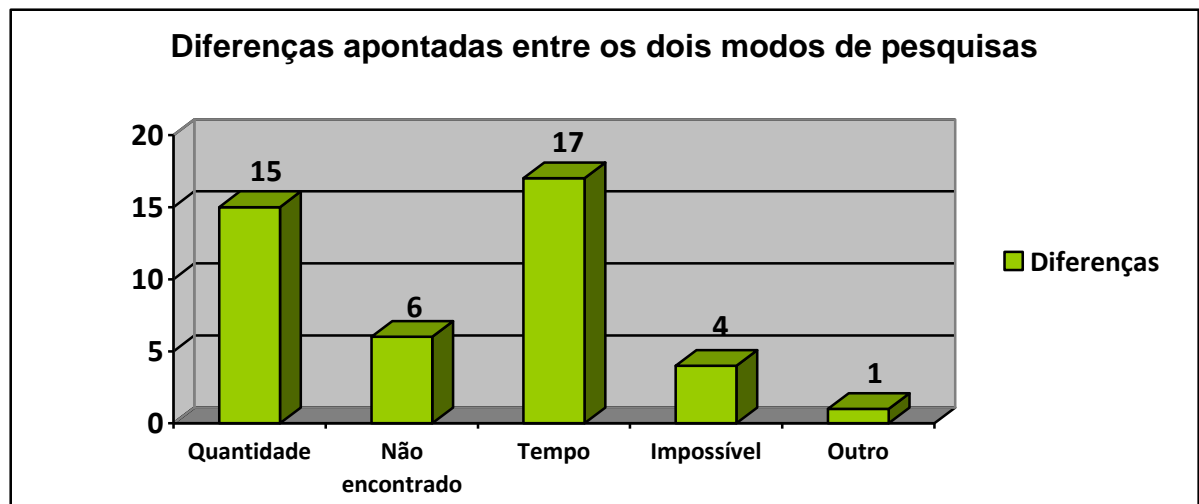


Gráfico 17 – Diferenças entre os dois modos de busca

Para comparação com o método utilizado para a busca, foi perguntado novamente ao entrevistado quantas tentativas agora, havia procedido para encontrar o produto no *site*.

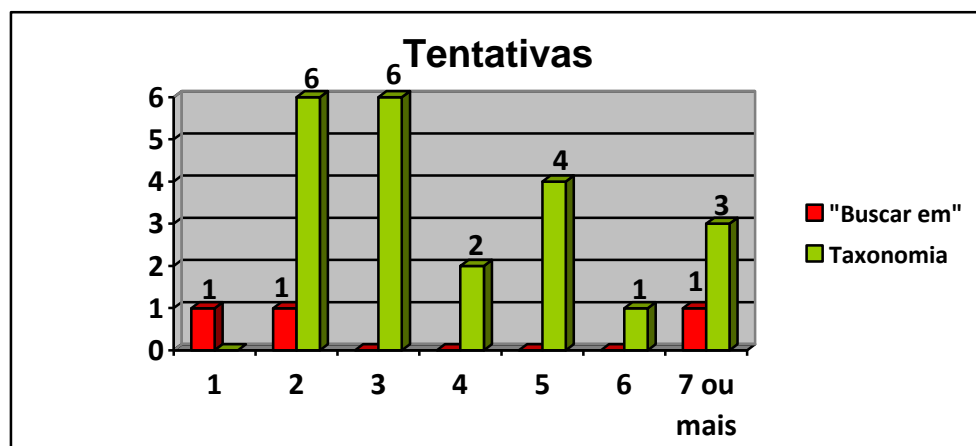


Gráfico 18 – Quantidade de tentativas para a segunda pesquisa

Em relação à busca livre, que permite maior expressão do usuário durante a pesquisa, 70% dos entrevistados não utilizaram nenhum instrumento para estratégia de busca (gráfico 17). Dos sujeitos da pesquisa que fizeram uso de algum desses instrumentos, ao total 9 participantes, responderam que foi possível encontrar o produto utilizando alguma estratégia (gráfico 18).

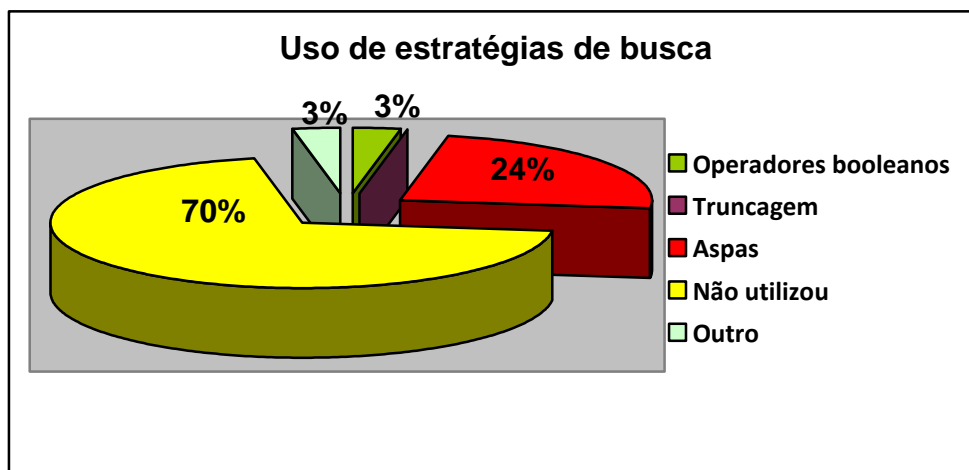


Gráfico 19 – Percentagem dos usuários que utilizaram algum instrumento para a busca

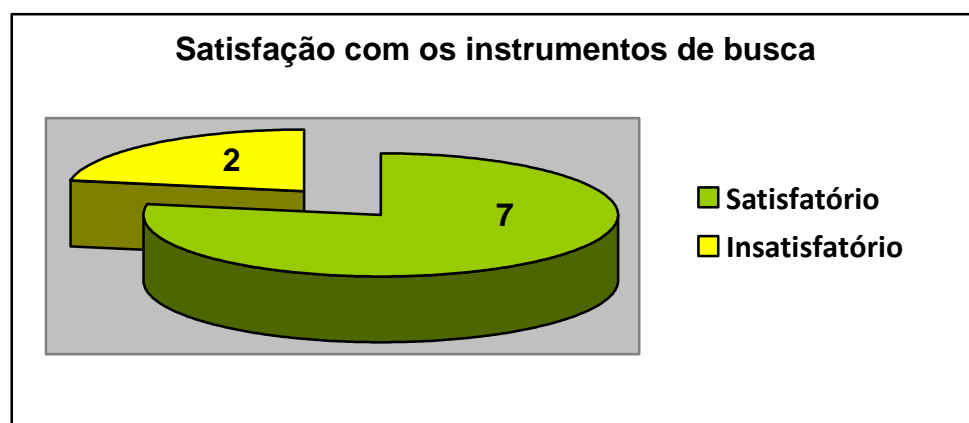


Gráfico 20 – Grau de satisfação quanto os resultados da pesquisa utilizando instrumentos de busca

7.5 Sobre a Livraria Cultura

As questões apresentadas para o estudo com a livraria Saraiva são iguais as que agora se referem à livraria Cultura, contudo as respostas apresentaram-se distintas em relação ao primeiro *site*.

Para a primeira pesquisa no *site* da Cultura, todos os entrevistados utilizaram inicialmente o campo “busca”, 100%, ou seja, os 30 participantes da pesquisa.

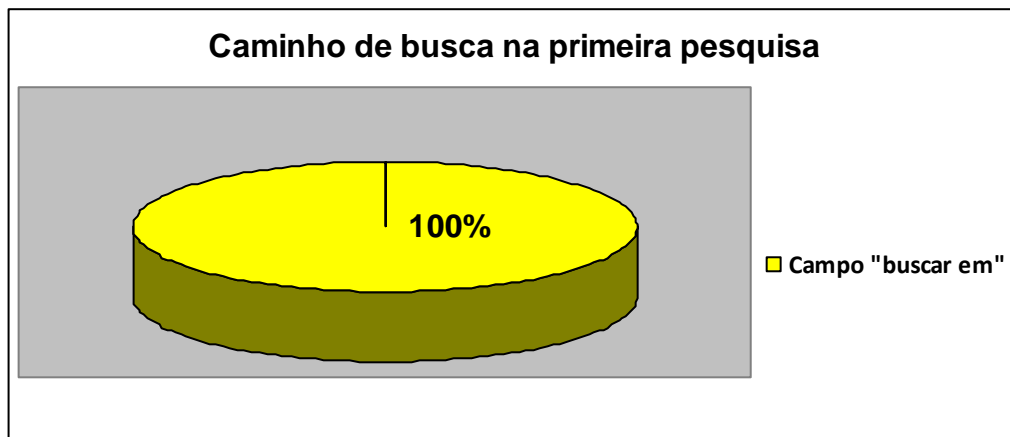


Gráfico 21 – Caminho utilizado pelos entrevistados para a primeira pesquisa

Com relação ao grau de satisfação quanto aos resultados apresentados nessa primeira busca, a maioria dos entrevistados, 77%, que utilizaram a busca livre respondeu que o produto foi recuperado satisfatoriamente, enquanto aqueles que não recuperaram o documento somam 23%.

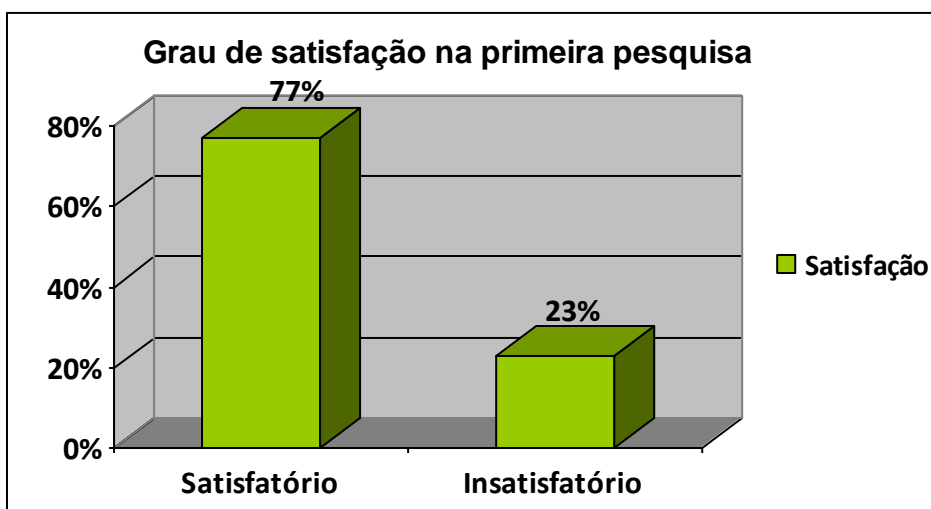


Gráfico 22 – Grau de satisfação na primeira pesquisa

A maior dificuldade apontada pelos entrevistados foi a falta de clareza do *site* e entre outras dificuldades citadas.

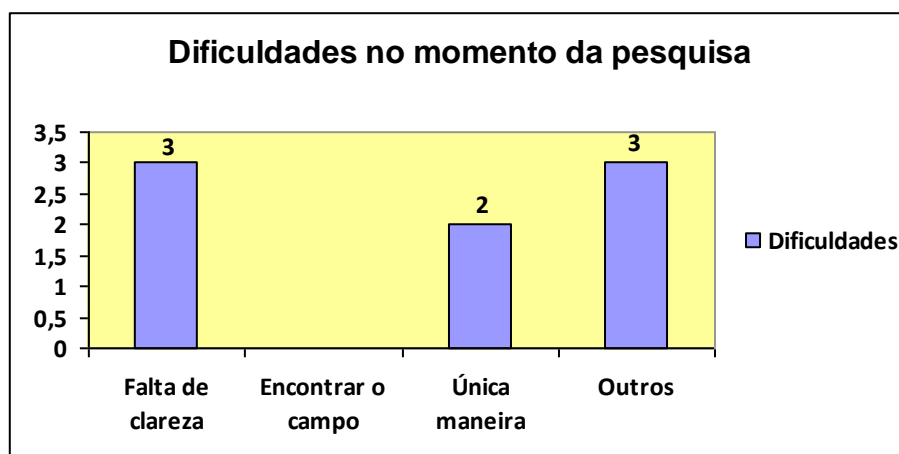


Gráfico 23 – Dificuldades encontradas no momento da pesquisa

Ao fim da primeira pesquisa, era necessário saber quantas tentativas os entrevistados realizaram para encontrar o produto.

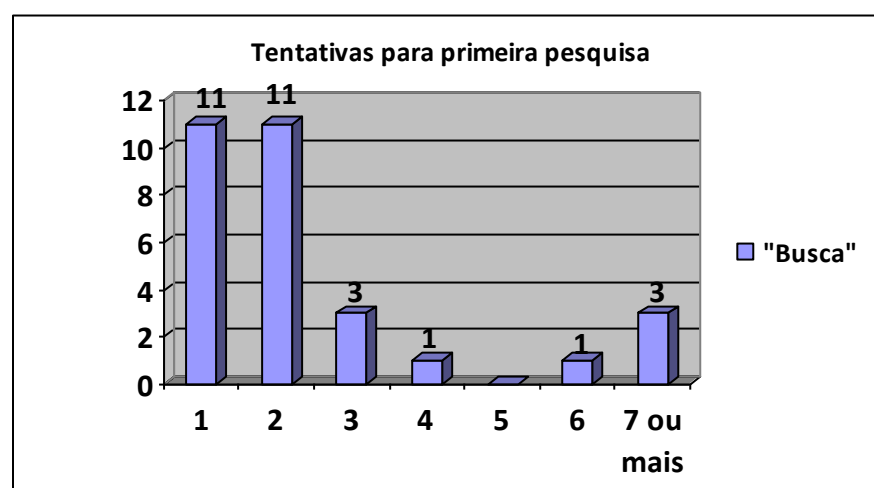


Gráfico 24 – Quantidade de tentativas para a primeira pesquisa

Com relação à segunda busca no *site* da livraria Cultura, 6 entrevistados responderam que esta busca apresentou resultados iguais a primeira, contudo 24 responderam que os produtos apresentados tinham sido diferentes, sendo que esta diferença era em relação ao tempo da pesquisa (18 citações) e o produto específico não havia sido encontrado (12 referências), com maiores representações na questão.

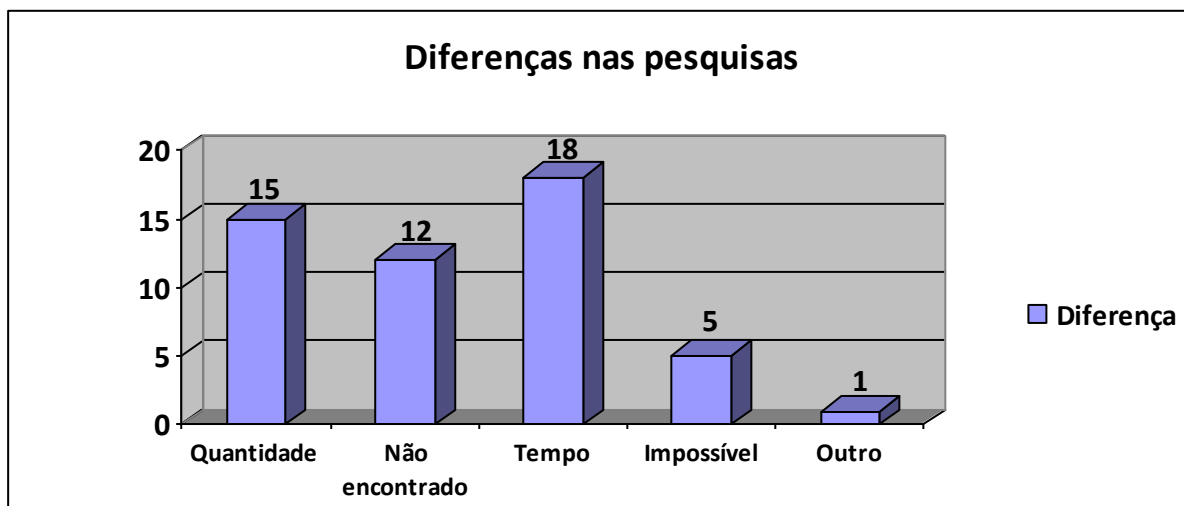


Gráfico 25 – Diferenças entre os dois modos de busca no *site* da Cultura

Como na primeira pesquisa todos os entrevistados utilizaram a busca livre, esta segunda pesquisa é referente à busca por meio da taxonomia. Do total dos entrevistados 10 responderam que precisaram fazer mais de 7 tentativas para encontrar o livro solicitado, por meio da pesquisa pela taxonomia.

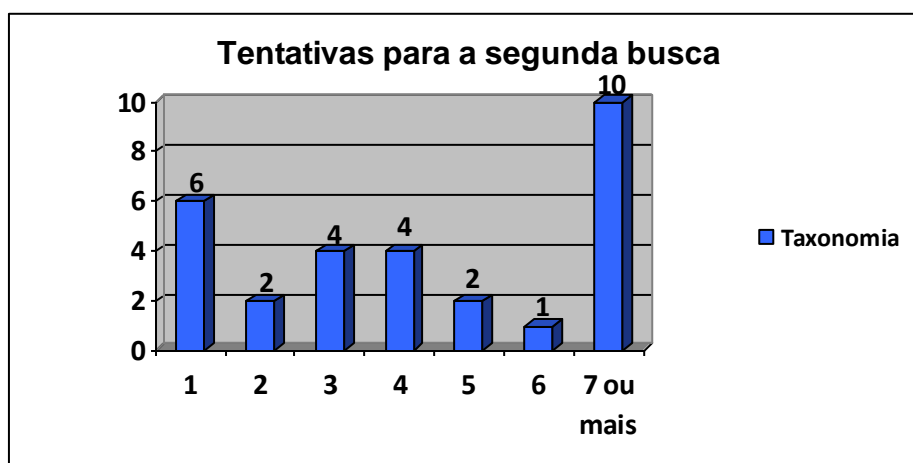


Gráfico 26 – Quantidade de tentativas para a segunda busca no *site* Cultura

Em relação à busca livre, apenas 7% fizeram uso de operadores booleanos na pesquisa, sendo que a grande maioria não utilizou nenhum instrumento para a pesquisa (80%), como apresenta o gráfico 26. Os entrevistados que utilizaram algum dos instrumentos no momento da busca, 67%, responderam que o resultado fora insatisfatório (gráfico 27).

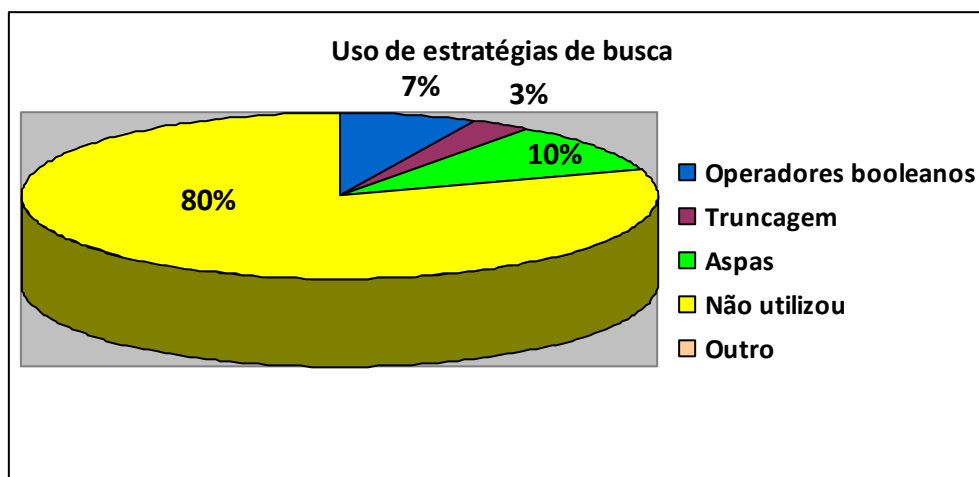


Gráfico 27 – Uso de estratégias de busca na busca livre

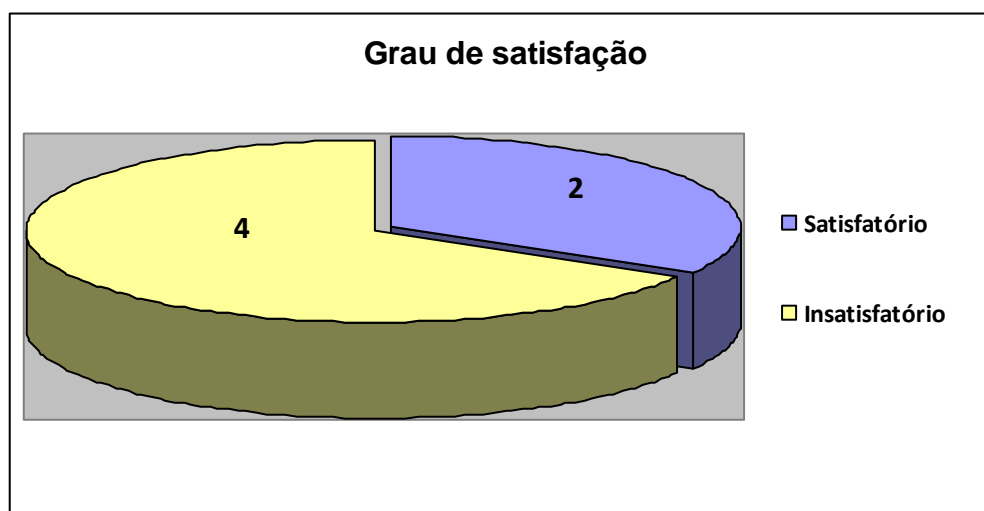


Gráfico 28 – Grau de satisfação com uso dos instrumentos de pesquisa

7.6 Busca livre x Busca controlada

Para avaliação final das duas livrarias, Saraiva e Cultura, e quanto ao modo de recuperação da informação, foram estabelecidos dois critérios na qual a análise está baseada. Os critérios escolhidos para a análise foram: tempo e os resultados apresentados pela pesquisa.

No *site* da livraria Saraiva, 83% do total dos entrevistados responderam que em relação ao tempo e aos resultados da pesquisa, a busca livre é a melhor opção para realizar uma pesquisa em *sites* como os das livrarias citadas, enquanto 17%

afirmaram que a taxonomia ou busca controlada é a melhor escolha. Para o *site* da livraria Cultura, 80% dos entrevistados avaliaram como o melhor método também a busca livre, enquanto que a taxonomia foi avaliada com 20%.

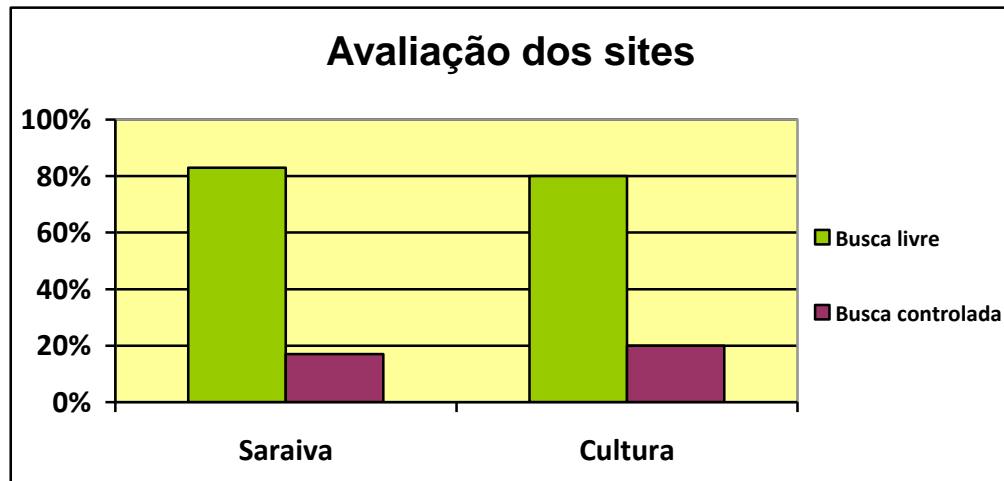


Gráfico 29 – Avaliação quanto os critérios tempo e resultados da pesquisa

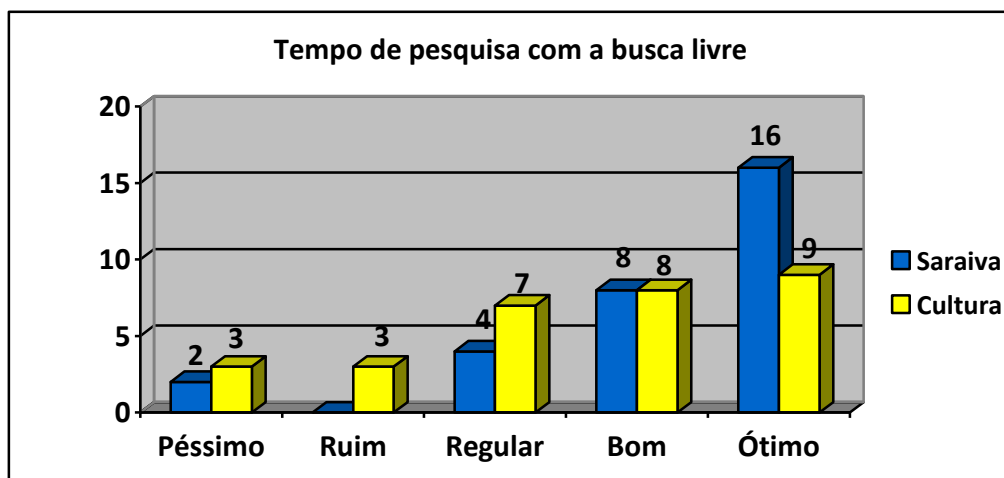


Gráfico 30 – Avaliação dos entrevistados quanto ao tempo da pesquisa

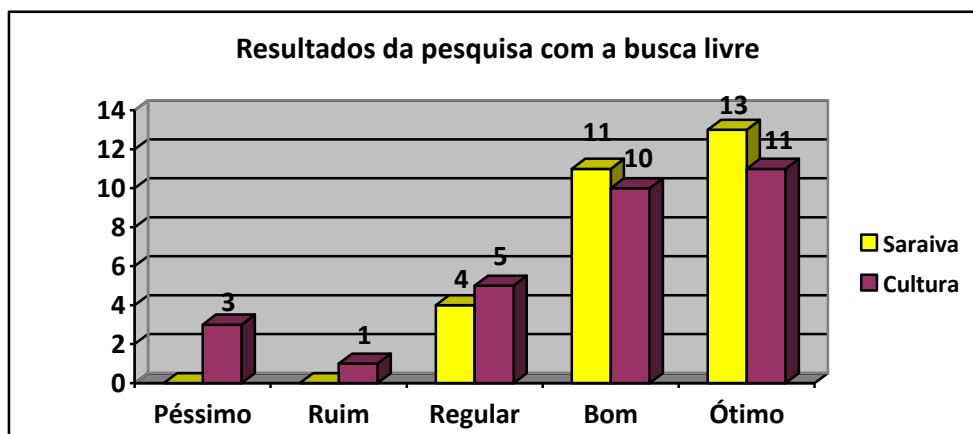


Gráfico 31 - Avaliação dos entrevistados quanto aos resultados da pesquisa

De acordo com as respostas dos entrevistados em relação à busca controlada há diferenças significativas em sua avaliação.

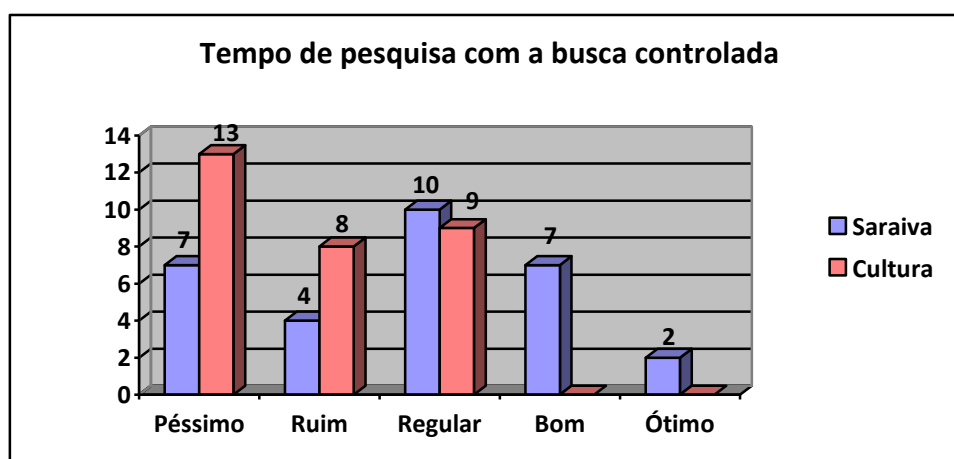


Gráfico 32 - Avaliação dos entrevistados quanto ao tempo da pesquisa

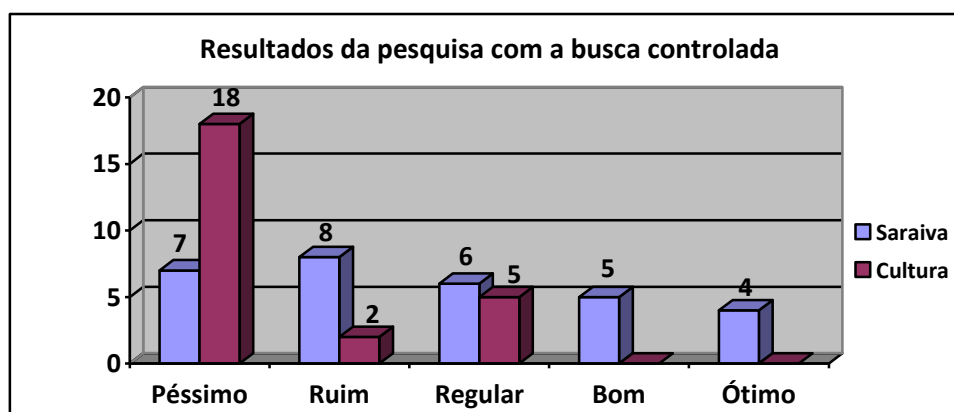


Gráfico 33 - Avaliação dos entrevistados quanto aos resultados da pesquisa

8 DISCUSSÃO DOS DADOS

As observações e análises apresentadas a seguir têm por base os dados expostos por meio dos gráficos neste trabalho e todos os outros resultados dos questionários.

A presença de estudantes em maior número neste estudo possibilita que as primeiras inferências¹⁰ sejam feitas a partir destes usuários. Os sujeitos entrevistados que se dizem estudantes preferem utilizar o campo de busca livre, na qual é possível fazer uso das estratégias de busca. Estas, contudo, também são as mais utilizadas por estes usuários. Os estudantes presentes na pesquisa são, em grande maioria, da graduação em Biblioteconomia da Universidade de Brasília, e que, portanto, já durante o curso estudam as estratégias de busca e com isso, apresentam maior facilidade ao realizar uma pesquisa.

Os entrevistados mais jovens, que compreendem a faixa etária de 15 a 25 anos de idade, demonstraram que preferem realizar uma busca livre na qual é possível que sejam feitas combinações para se encontrar a informação que precisa. A razão para este caso pode ter relação ao fato de nessa faixa etária serem na maioria estudantes e também, por estarem frequentemente fazendo uso da internet e pesquisas na web.

Para a primeira busca, a qual deveria ser realizada de acordo com a vontade de cada sujeito participante, há a preferência na grande maioria pela busca livre, que pode ser explicada ainda, de acordo com a observação realizada durante a aplicação do questionário, como uma forma mais rápida para se chegar ao produto. No roteiro que seguia junto ao questionário havia dados que possibilitavam ao usuário utilizar diferentes termos para a pesquisa, tais como o nome do autor e ano.

O título do livro utilizado na pesquisa causou dificuldades a muitos entrevistados, exceto aos estudantes de Biblioteconomia que já tinham conhecimento que o livro era da sua área de graduação. Aqueles que não sabiam que se tratava de um livro da Ciência da Informação deduziram pelo título do mesmo, que estava ligado à área da Pedagogia, Didáticos ou ainda a Manuais. Esse

¹⁰ Inferência: Processo mental que por meio de dados chega-se a uma conclusão.

fato dificultou principalmente a pesquisa por meio da taxonomia, pois os usuários deveriam “descobrir” em qual categoria encontrava o livro.

Houve maior dificuldade em relação ao *site* da livraria Cultura, pois o livro não se encontra na lista oferecida pela taxonomia, ainda que esteja classificado na área de Pedagogia. Por esse motivo, houve o grande número de tentativas utilizando a taxonomia, o que levou ao depoimento de um entrevistado frustrado quanto ao resultado apresentado pelo *site*: “a busca controlada foi horrível. Não consegui encontrar o livro em lugar nenhum!” Esse fato pode desmotivar o usuário a utilizar o *site* da livraria, buscando assim outros que melhor apresentem os produtos.

Outro fato, em relação à livraria Cultura, é que os livros de Biblioteconomia se encontram na subdivisão da área da Comunicação, impossibilitando assim, até mesmo a um estudante de Biblioteconomia, encontrar o livro solicitado na pesquisa. Por um raciocínio simples da grande maioria, não haveria ligação direta entre a Comunicação e a Ciência da Informação. Por outro lado, no *site* da Saraiva, a seção com os livros de Biblioteconomia encontram-se disponíveis na área de Ciências Sociais e Aplicadas, sendo, portanto, melhor avaliada pelos entrevistados, como neste comentário, referente a questão aberta do item 6.2: “gostei da forma como é (sic) dividida as categorias no *site* da Saraiva, o termo biblioteconomia está bem posicionado – dentro de ciências humanas e sociais, portanto fácil de ser localizado”.

De acordo com as taxonomias dos dois *sites*, aparentemente percebemos que não houve um tratamento adequado em relação à categorização e indexação dos elementos, pois ao se realizar a busca por um livro na área da Biblioteconomia, documentos relacionados às outras áreas, como por exemplo, Arquivologia, também são recuperados. Provavelmente isso ocorre porque na grande maioria não há um profissional da área, bibliotecário, para que auxilie e execute essa tarefa. Esse fato pode explicar o que aconteceu durante a pesquisa: os entrevistados que pesquisaram primeiramente pela busca livre recuperaram menos documentos do que na segunda pesquisa, utilizando a taxonomia, pois esta recuperou documentos além da área da Biblioteconomia, como demonstra o comentário de outro participante: “há exposição de grande número de obras que não têm afinidade com o assunto especificado” (questão 9.2, referente à livraria Cultura).

A busca livre também apresentou dificuldades aos entrevistados, uma vez que não poderiam pesquisar pelo título completo do livro, caso não houvesse essa

informação, a pesquisa ficaria ainda mais complicada. Alguns deles optaram por utilizar a informação do nome do autor contida no roteiro para realizar a busca. Outra forma de pesquisa foi a utilização de palavras do título, tais como biblioteca, escola. Alguns entrevistados tentaram utilizar os operadores booleanos ou a técnica de truncagem, porém, como apresentado na revisão de literatura deste trabalho, nenhum dos dois *sites* permite o seu uso, e ao tentar utilizar, o *site* remete à página de pesquisa avançada. Esse último modo de pesquisa, também não apresentou resultados satisfatórios aos usuários, como pode ser comprovado em um comentário na questão aberta do questionário: “o sistema de busca avançada apresentado no *site* dificulta a pesquisa, consequentemente a localização do livro de interesse” (resposta ao item 9.2).

A avaliação final dos entrevistados reflete bem a dificuldade encontrada ao utilizar a linguagem controlada, pois preferiram, na grande maioria pelo uso da busca livre. A preferência por este tipo de busca pode estar diretamente influenciada pelo uso frequente, principalmente por entrevistados que comumente realizam pesquisas na internet, do mecanismo de busca atual, o *Google*. A forma de busca dessa ferramenta, que apresenta janela para o usuário realizar a busca livre tornou-se praticamente um modelo de busca na internet. Os usuários estão familiarizados com esse modo de pesquisa livre, mais conhecido que as taxonomias.

As dificuldades com o uso de taxonomias estão relacionadas também ao fato de duas pessoas não estruturarem os assuntos da mesma maneira. Cada pessoa fará de acordo com o que acredita ser mais conveniente, o que traz problemas com a padronização da organização das informações.

A maioria dos comentários referentes às questões abertas dos itens 6.2 e 9.2 do questionário reflete bem a dificuldade encontrada pelos usuários ao realizar a pesquisa pela taxonomia: “achei péssima a categorização dos assuntos. Simplesmente não encontrei biblioteconomia ou ciência da informação. Se não tivesse a busca livre, seria bem difícil achar a obra” (questão a respeito da Livraria Cultura).

Há ainda aqueles que demonstraram frustração a ambos os processos de busca nos *sites*, como o comentário do entrevistado: “para ambos, os processos são demorados. É melhor conversar com o vendedor”. O tempo de pesquisa perdido, principalmente com o uso da taxonomia, pode causar ao entrevistado ainda mais aversão à taxonomia.

9 CONSIDERAÇÕES FINAIS

A dificuldade de realizar uma pesquisa na internet por algum documento exige tempo, esforço e habilidade para que a recuperação da informação seja eficiente e corresponda ao desejo do usuário. O estudo de caso com as livrarias eletrônicas permitiu a avaliação quanto ao uso das duas linguagens: natural e controlada.

Como resposta ao problema deste estudo, a adoção das duas linguagens apresentou, de maneira geral, resultados satisfatórios aos participantes da pesquisa. Porém, a busca livre, como já analisada anteriormente, apresentou resultados mais satisfatórios quanto ao tempo e aos resultados da pesquisa do que a taxonomia. Contudo, vale lembrar que a disponibilização das duas linguagens é mais recomendada na literatura (Lopes, 2002 e Lancaster, 2004). O uso de apenas uma delas nos *sites* pode limitar a forma de recuperação da informação, e obrigar o usuário a usar uma única forma de pesquisa pelo produto.

É importante também salientar que as duas linguagens devem estar bem elaboradas para que não ocorram divergências no resultado da pesquisa ao usuário final.

9.2 Limitações da pesquisa e sugestões para estudos futuros

A busca por apenas um produto nos dois *sites* pode não ter possibilitado a generalização dos resultados em relação a outros itens das livrarias eletrônicas escolhidas para o estudo de caso neste trabalho. Porém, a utilização de mais de um produto para que o entrevistado realizasse a pesquisa iria aumentar o questionário, e conseqüentemente, dificultaria a aplicação do estudo.

O questionário também merece destaque, pois inicialmente a proposta era aplicar a pesquisa em três *sites* de livrarias eletrônicas, que seriam a livraria Saraiva, a Cultura e a Siciliano. A Saraiva e a Siciliano têm um convênio firmado entre as suas lojas na internet, e por isso seus *sites* apresentam ferramentas semelhantes para a busca. Optamos, portanto, aplicar o estudo na Livraria Saraiva e Cultura.

Outra dificuldade encontrada foi adequar o vocabulário técnico do roteiro e do questionário à linguagem do usuário, utilizando termos mais coloquiais, como por exemplo, não utilizar a palavra taxonomia, mas sim categorias ou listas de assuntos. A adoção desses termos mais comumente utilizados pelos entrevistados visava a não criar dúvidas para responder as questões. Mas ainda sim, com o cuidado tomado, algumas dúvidas surgiram e que foram respondidas pelo entrevistador no momento da pesquisa ou anteriormente a esta, após a leitura do roteiro pelo participante.

Apesar das dificuldades, a importância do trabalho foi reconhecida por alguns entrevistados. Foi possível observar que aqueles que fizeram comentários positivos após a aplicação da pesquisa se interessaram pelo assunto e pediram maiores explicações sobre alguns itens abordados, como por exemplo, os operadores booleanos, o que são e como utilizá-los na pesquisa.

Como sugestão para trabalhos futuros, sugere-se que a avaliação das linguagens natural e controlada possam ser estudadas em áreas diferentes e que são de uso comuns das pessoas, para que as formas de buscas sejam mais adequadas às necessidades dos usuários. Também, sugere-se o uso de mais de um produto na pesquisa e que sejam de assuntos distintos para que se possam fazer mais generalizações com os resultados.

REFERÊNCIAS

ABDALA, Elisabeth Ávila; OLIVEIRA, Mirian. Formas de pagamento utilizadas pelas livrarias eletrônicas brasileiras. **Caderno de Pesquisas em Administração**, São Paulo, v. 09, nº 4, out./dez. 2002. Disponível em: <ancibe.com.br/.../artigo%20-%20Formas%20de%20pagamento/.../pdf>. Acesso em: 20 jul. 2010.

ARAÚJO JÚNIOR, Rogério Henrique de. **Precisão no processo de busca e recuperação da informação**. Brasília: Thesaurus, 2007. 171 p.

BAPTISTA, Sofia Galvão; CUNHA, Murilo Bastos da. Estudo de usuários: visão global dos métodos de coleta de dados. **Perspectivas em Ciência da Informação**, v. 12, n. 2, p. 168-184, maio/ago. 2007. Disponível em: <revista.ibict.br/pbcib/index.php/pbcib/.../702>. Acesso em: 15 jul. 2010.

BOCCATO, Vera Regina Casari.; FUJITA, Mariângela Spotti Lopes. **Avaliação de linguagem documentária em fonoaudiologia na perspectiva do usuário: estudo de observação da recuperação da informação com protocolo verbal**. Marília: UNESP, 2005. Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual Paulista. Disponível em: <www.marilia.unesp.br/.../CienciadaInformacao/.../boccato_vrc_me_mar.pdf>. Acesso em 03 jun. 2010.

BRANSKI, Regina Meyer. Recuperação de informações na web. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 9, n. 1, p. 70-87, jan./jun. 2004. Disponível em: <portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/351/160>. Acesso em: 13 maio 2010.

BRÄSCHER, Marisa. A ambigüidade na recuperação da informação. **Revista de Ciência da Informação**, v. 3, n. 1, fev. 2002. Disponível em: <repositorio.bce.unb.br/handle/10482/924>. Acesso em: 13 jan. 2010.

CAMPOS, Maria Luiza de Almeida. **Linguagem documentária**: teorias que fundamentam sua elaboração. Niterói: EdUFF, 2001. 133 p.

CARLAN, Eliana. **Introdução à elaboração de tesauro**. Brasília: UnB, [2008]. 50 slides: color.

CARLAN, Eliana. **Sistemas de organização do conhecimento**: uma reflexão no contexto da Ciência da Informação. Brasília: UnB, 2010. 195 p. Dissertação (Mestrado em Ciência da Informação) – Universidade de Brasília.

CAVALCANTI, Cordélia R. **Indexação e tesauro**: metodologia e técnicas. Brasília: ABDF, 1978. 89 p.

CEGALLA, Domingos Paschoal. **Novíssima gramática da língua portuguesa**. 46. ed. São Paulo: Companhia Editora Nacional, 2005. 693 p.

CINTRA et al. **Para entender as linguagens documentárias**. 2.ed. rev. e ampl. São Paulo: Polis, 2002. 92 p. cap. 2 e 4.

COIMBRA, Fabíola de Luca. **Avaliação da linguagem natural e controlada na recuperação de doutrina jurídica**. Brasília: UnB, 2005. 132 p. Dissertação (Mestrado em Ciência da Informação) – Universidade de Brasília.

DANTAS, Suellen Fernandes. **Recuperação da informação e avaliação de usabilidade dos webOPACS dos sistemas Aleph e Sophia**. Brasília: UnB, 2008. 112 p. Monografia (Graduação em Biblioteconomia) – Universidade de Brasília.

DODEBEI, Vera Lúcia Doyle. **Tesauro**: linguagem de representação da memória documentária. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002. 119 p.

DUQUE, Cláudio Gottschalg-. **Sirilico**: uma proposta para um sistema de recuperação de informação baseado em teorias da lingüística computacional e ontologia. Belo Horizonte: UFMG, 2005. 120 p. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais.

FIDEL, Raya. Searchers' selection of search Keys: II. Controlled vocabulary or free-text searching. **Journal of the American Society for Information Science**, v. 42, n. 7, 1991. Disponível em: <citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.1501.pdf>. Acesso em: 02 fev. 2010.

GASPERIN, Caroline Varaschin; LIMA, Vera Lúcia Strube de. Fundamentos do processamento estatístico da linguagem natural, **Relatório Técnico**, n. 21, 2001. Disponível em: <>. Acesso em: 11 ago. 2010

GOMES, Hagar Espanha. Classificação, tesouro e terminologia: fundamentos comuns. 1996. Disponível em: <www.conexaorio.com/bit/tertulia/tertulia.htm>. Acesso em: 03 fev. 2010.

HÜBNER, M. Martha. **Guia para elaboração de monografias e projetos de dissertação e doutorado**. São Paulo: Pioneira/Mackenzie, 1998.

LANCASTER, F. W. **Indexação e resumos: teoria e prática**. 2. e.d. Brasília: Briquet de Lemos, 2003. 452 p.

LEISE, Fred; FAST, Karl; STECKEL, Mike. What is a controlled vocabulary? December, 2002. Disponível em: <www.bboxesandarrows.com/view/what_is_a_controlled_vocabulary_>. Acesso em: 07 maio 2010.

Lobiondo-Wood G, Haber J. **Pesquisa em enfermagem: métodos, avaliação crítica e utilização**. Rio de Janeiro: Guanabara Koogan, 2001.

LOPES, Ilza Leite. **Curso de indexação**. Brasília: UnB, [2007]. 43 slides: color.

LOPES, Ilza Leite. Estratégias de busca na recuperação da informação: revisão de literatura. **Revista de Ciência da Informação**, Brasília, v. 31, n. 2, p. 60-71,

maio/ago. 2002. Disponível em: <<http://www.scielo.br/pdf/ci/v31n2/12909.pdf>>. Acesso em: 07 jan. 2010.

LOPES, Ilza Leite. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Revista da Ciência da Informação**, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002. Disponível em: <www.scielo.br/pdf/ci/v31n1/a05v31n1.pdf>. Acesso em: 13 jan. 2010.

MORELLATO, Luana. **Processamento de linguagem natural**. Disponível em: <http://imasters.uol.com.br/artigo/8537/tendencias/processamento_de_linguagem_natural/>. Acesso em: 11 ago. 2010.

MOURA, Gevilacio Aguiar Coêlho de. Sistemas de busca da web: diretórios e mecanismos de busca. [online] Disponível em: <www.quatrocantos.com/tec_web/sist_busca/index.htm>. Acesso em: 12 jun. 2010.

MUDDAMALLE, Manikya Rao. Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. **Journal of the American Society for Information Science**, v. 49, n. 10, 1998. Disponível em: <nlp.korea.ac.kr/.../%5BMuddamalle98%5Dnaturallanguagevccontrolledvocinir.pdf>. Acesso em: 02 fev. 2010.

NACHMIAS, Rafi. Needle in a hyperstack: searching information on the world wide web. **Journal of research on computing in education**, March, 2002. Disponível em: <<http://muse.tau.ac.il/publications/70.pdf>>. Acesso em: 04 maio 2010.

OLIVEIRA, Fábio Abreu Dias de. Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa. 2009. Disponível em: <www.inf.ufrgs.br/gppd/disc/cmp135/.../parser.html>. Acesso em: 10 ago. 2010.

OTHERO, Gabriel de Ávila. Linguística computacional: uma breve introdução. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 341-351, jun. 2006. Disponível em:

<<http://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/viewFile/605/436>>.

Acesso em: 10 ago. 2010.

SAYÃO, Luís Fernando. Bases de dados: a metáfora da memória científica. **Ciência da Informação**, Brasília, v.25, n.3, 1996.

TERRA, José Cláudio C. et. al. Taxonomia: elemento fundamental para a gestão do conhecimento. 2005. Disponível em:

<www.terraforum.com.br/.../taxonomia_%20fundamental_GC.pdf>. Acesso em: 22 abr. 2010.

VOGEL, Michely Jabala M. Taxonomia: alguns conceitos e algumas confusões. 2009. Disponível:

<www.terraforum.com.br/.../Taxonomia_%20conceitos_confusoes.pdf>. Acesso em: 22 abr. 2010.

YAMAOKA, Eloi Juniti. Recuperação de informação na web. Brasília: 2003. Disponível em: < >. Acesso em: 02 maio 2010.

ANEXO I – Respostas da questão aberta 6.2 referente à Livraria Saraiva.

1. Em relação a livraria Saraiva, acredito que o site deveria organizar melhor os títulos, pois na Livraria Cultura foi bem mais fácil achar o livro informado. Quanto ao estudo, acho bem interessante, mais legal seria se desse dicas de como fazer uma pesquisa mais detalhada e clara.
2. As hierarquias são muito mal estruturadas. Por exemplo, poderia ter um + do lado do termo ciências sociais aplicadas, para indicar que existem subitens.
3. Na busca controlada a editora deveria diversificar, exemplo: livros: -didática, científico, etc.
4. O autor poderia ser identificado por suas obras com a opinião para especificar o assunto.
5. Para ambos, os processos são demorados. É melhor conversar com o vendedor.
6. Creio que tendo o pesquisador conhecimento do que procura, qualquer que seja o método utilizado terá resultado satisfatório.
7. Gostei da forma como é dividida as categorias no site da Saraiva, o termo *biblioteconomia* está bem posicionado – dentro de ciências humanas e sociais, portanto fácil de ser localizado.
8. O sistema de busca avançada apresentado no site dificulta a pesquisa e consequentemente a localização do livro de interesse.
9. Acho um pouco fraco para pesquisas para quem não sabe todas as especificações do livro que esta buscando.
10. Esta foi a mais complicada de pesquisar. Possui várias ordens hierárquicas, dificultando a busca e aumentando o tempo da pesquisa.

ANEXO II – Respostas da questão aberta 9.2 referente à Livraria Cultura.

1. A busca controlada foi horrível. Não consegui encontrar o livro em lugar nenhum!
2. A Livraria Cultura está de parabéns, pois apresenta uma filtragem básica, você consegue achar o que deseja sem ter que tentar mais de uma vez. O estudo tem a finalidade de analisar e pesquisar os meios de pesquisa nos sites de livros, acho bem bacana. Na internet o que vale mesmo é a praticidade. Às vezes você demora tanto para achar o que quer em alguns sites, que é melhor ir pessoalmente na loja.
3. O site não apresenta resultados claros na busca controlada.
4. A pesquisa utilizando o 2º caminho é pouco eficiente e demanda maior tempo.
5. Há exposição de grande número de obras que não têm afinidade com o assunto especificado.
6. O cliente tem que ter paciência.
7. Achei péssima a categorização dos assuntos. Simplesmente não encontrei biblioteconomia ou ciência da informação. Se não tivesse a busca livre, seria bem difícil achar a obra.
8. O sistema de busca avançada apresentado no site dificulta a pesquisa e consequentemente a localização do livro de interesse.
9. Já tinha uma certa experiência na utilização do site da cultura o que acabou facilitando a busca.

APÊNDICE I – Roteiro com orientações aos usuários para a realização da pesquisa

ROTEIRO

Desde já, gostaria de agradecer a sua participação e dizer que é fundamental que este questionário seja respondido seriamente. São apenas necessários alguns minutos para a realização da atividade, da qual as respostas serão utilizadas para análise do estudo de caso.

Seguindo a ordem do questionário, as questões de 1 a 3 e respectivamente seus subitens, referem-se ao perfil do entrevistado, com o objetivo de identificar o contexto do sujeito.

As questões que seguem, do item 4 ao 12 e respectivamente seus subitens, deverão ser respondidos após a navegação nos *sites* das livrarias propostas para este estudo e seguindo as orientações a seguir.

- Livraria Saraiva: www.saraiva.com.br
- Livraria Cultura: www.livrariacultura.com.br

Para a realização da atividade com as livrarias eletrônicas são propostas duas etapas: **1ª etapa:** O entrevistado deverá a sua maneira pesquisar pelo produto sugerido. **2ª etapa:** O entrevistado deverá fazer uma segunda pesquisa, porém utilizando o modo de busca sugerido pela pesquisa.

- **Busca livre:** utiliza-se o campo *busca* (no *site* da Cultura) ou *buscar em* (*site* da Saraiva);
- **Busca controlada:** utiliza-se a lista de assuntos categorizados e hierarquizados pelo próprio *site*.

Produto definido para a pesquisa

O entrevistado deverá realizar a busca nos dois *sites* pelo **livro**:

Título: Como usar a biblioteca na escola	Autor: Carol Kuhlthau
Ano: 2002	Assunto: Biblioteconomia
Editora: Autentica-Pedagogia	

➤ Uma regra:

- Para atingir os objetivos da pesquisa, a busca livre não deve ser feita pelo título do produto.

Obrigada.
Tainá Batista de Assis
Estudante de Biblioteconomia – UnB
Monografia: O uso da linguagem natural e controlada na recuperação da informação na web: o caso das livrarias eletrônicas.

APÊNDICE II – Questionário de avaliação dos *sítes* das livrarias Saraiva e Cultura.

1. Perfil do entrevistado

1.1 Idade: anos

1.2 Sexo: () F () M

1.3 Profissão:

1.4 Com qual frequência faz uso da internet:

() Frequentemente () Regularmente () Raramente () Nunca

1.5 Com qual frequência faz pesquisas na Internet:

() Frequentemente () Regularmente () Raramente () Nunca

2. Conhecimento sobre estratégias de busca

2.1 Ao realizar uma pesquisa na internet, você conhece e/ou faz uso de algum instrumento para refinar a busca? Qual? **(Se necessário, marque mais de uma opção).**

() Sim, conheço e/ou utilizo os operadores booleanos (AND, OR, NOT)

() Sim, conheço e/ou utilizo a técnica de truncagem.

() Sim, conheço e/ou utilizo as aspas.

() Conheço, mas não utilizo.

() Não conheço e nem utilizo nenhum instrumento.

() Outro: _____

3. Conhecimento dos *sites* da Livrarias Saraiva, Cultura e Siciliano

3.1 Já realizou alguma busca por produtos nos *sites* das livrarias Saraiva, Cultura e/ou Siciliano? **(Se necessário, marque mais de uma opção).**

() Sim, já fiz pesquisa em todos.

() Somente pesquisei no *site* da Livraria Saraiva.

() Somente pesquisei no *site* da Livraria Cultura.

() Somente pesquisei no *site* da Livraria Siciliano.

() Não pesquisei em nenhum desses *sites*.

4 Sobre a Livraria Saraiva

4.1 Para a primeira busca no *site*, qual o caminho você utilizou?

() Fiz a busca utilizando o campo “buscar em” e digitei o que precisa.

☐ Fiz a busca percorrendo os assuntos hierarquizados pelo próprio *site*.

() Outro. Qual? _____

4.2 O resultado da busca pelo produto (previamente estabelecido para este estudo) foi satisfatório?

() Sim, tudo o que pesquisei foi encontrado de forma satisfatória.

☐ Não, o que pesquisei não foi recuperado.

4.3 Você encontrou alguma dificuldade para fazer a busca?

() Sim () Não

4.4 Caso sim, sua dificuldade foi em relação a: **(Se necessário, marque mais de uma opção)**

- ☐) Encontrar no *site* o campo para iniciar a pesquisa.
- ☐) Falta de clareza dos assuntos listados pelo *site*.
- ☐) O *site* apresenta apenas uma maneira para pesquisar.
- ☐) Outros: _____.

4.5 Quantas tentativas você realizou para encontrar o produto solicitado?

- ☐) 1 ☐) 2 ☐) 3 ☐) 4 ☐) 5 ☐) 6 ☐) 7 ou mais

4.6 Agora, realizada a segunda busca (conforme explicado no roteiro), o caminho percorrido, diferentemente do primeiro, apresentou os mesmos resultados comparando com o de antes?

- ☐) Não, o resultado apresentado foi totalmente diferente que o primeiro.
- ☐) Sim, o resultado apresentado foi igual ao primeiro.

4.7 Caso não, o que apresentou de diferente? **(Se necessário, marque mais de uma opção).**

- ☐) Na quantidade de produtos recuperados.
- ☐) O produto específico pesquisado não foi encontrado.
- ☐) O tempo de pesquisa.
- ☐) Por este caminho não foi possível realizar a pesquisa.
- ☐) Outros: _____.

4.8 Quantos caminhos você percorreu para encontrar o produto solicitado?

- ☐) 1 ☐) 2 ☐) 3 ☐) 4 ☐) 5 ☐) 6 ☐) 7 ou mais

4.9 Na busca livre, no campo “buscar em” no *site* fez uso ou tentou usar alguma estratégia de busca? **(Se necessário, marque mais de uma opção).**

- ☐) Sim, utilizei ou tentei os operadores booleanos (AND, OR, NOT)
- ☐) Sim, utilizei ou tentei a técnica de truncagem.
- ☐) Sim, utilizei ou tentei as aspas.
- ☐) Não utilizei nem tentei algum instrumento.
- ☐) Outro: _____.

4.10 Caso tenha utilizado, o resultado foi satisfatório?

- ☐) Sim, encontrei o produto.
- ☐) Não, não foi possível realizar a pesquisa.

5 Quanto aos critérios de tempo e resultados da pesquisa, considere:

1 Péssimo 2 Ruim 3 Regular 4 Bom 5 Ótimo

5.1 A busca livre em relação ao tempo:

- ☐) 1 ☐) 2 ☐) 3 ☐) 4 ☐) 5

5.2 A busca livre em relação aos resultados:

- ☐) 1 ☐) 2 ☐) 3 ☐) 4 ☐) 5

5.3 A busca controlada em relação ao tempo:

- ☐) 1 ☐) 2 ☐) 3 ☐) 4 ☐) 5

5.4 A busca controlada em relação aos resultados:

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

6 Considerações finais para o estudo na Livraria Saraiva

6.1 Após realizado os dois tipos de pesquisas e comparando-os, qual apresentou os resultados mais satisfatórios?

- ☐ A busca livre, na qual digitei o que precisava.
☐ Utilizando a lista de assuntos definidos pelo *site*.

6.2 Desejaria fazer algum comentário em relação a este estudo ou sobre a pesquisa no *site* da Livraria Saraiva?

7 Sobre a Livraria Cultura

7.1 Para a primeira busca no *site*, qual o caminho você utilizou?

- ☐ Fiz a busca utilizando o campo “busca” e digitei o que precisa.
☐ Fiz a busca percorrendo os assuntos hierarquizados pelo próprio *site*.
☐ Outro. Qual? _____.

7.2 A busca pelo produto (previamente estabelecido para este estudo) apresentou resultados satisfatórios?

- ☐ Sim, o que pesquisei foi encontrado de forma satisfatória.
☐ Não, o que pesquisei não foi recuperado.

7.3 Você encontrou alguma dificuldade para fazer a busca?

- ☐ Sim ☐ Não

7.4 Caso sim, sua dificuldade foi em relação a: **(Se necessário, marque mais de uma opção).**

- ☐ Encontrar no *site* o campo para iniciar a pesquisa.
☐ Falta de clareza dos assuntos listados pelo *site*.
☐ O *site* apresenta apenas uma maneira de pesquisar.
☐ Outros: _____.

7.5 Quantas tentativas você realizou para encontrar o produto solicitado?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ou mais

7.6 Agora, realizada a segunda busca (conforme explicado no roteiro), o caminho percorrido, diferentemente do primeiro, apresentou os mesmos resultados comparando com o de antes?

- ☐ Não, o resultado apresentado foi totalmente diferente que o primeiro.
☐ Sim, o resultado apresentado foi igual ao primeiro.

7.7 Caso não, o que apresentou de diferente? (Se necessário, marque mais de uma opção).

- ☐ Na quantidade de produtos recuperados.
- ☐ O produto específico pesquisado não foi encontrado.
- ☐ O tempo de pesquisa.
- ☐ Com o caminho escolhido não foi possível realizar a pesquisa.
- ☐ Outros:_____.

7.8 Quantos caminhos você percorreu para encontrar o produto solicitado?

- ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ou mais

7.9 Na busca livre, no campo “buscar em” no *site* fez uso ou tentou usar alguma estratégia de busca? (Se necessário, marque mais de uma opção).

- ☐ Sim, utilizei ou tentei os operadores booleanos (AND, OR, NOT)
- ☐ Sim, utilizei ou tentei a técnica de truncagem.
- ☐ Sim, utilizei ou tentei as aspas.
- ☐ Não utilizei nem tentei algum instrumento.

7.10 Caso tenha utilizado, o resultado foi satisfatório?

- ☐ Sim, encontrei o produto.
- ☐ Não, não foi possível realizar a pesquisa.

8 Quanto aos critérios de tempo e resultados da pesquisa, considere:

1 Péssimo 2 Ruim 3 Regular 4 Bom 5 Ótimo

8.1 A busca livre em relação ao tempo:

- ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

8.2 A busca livre em relação aos resultados:

- ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

8.3 A busca controlada em relação ao tempo:

- ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

8.4 A busca controlada em relação aos resultados:

- ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

9 Considerações finais para o estudo na Livraria Cultura

9.1 Após realizado os dois tipos de pesquisas e comparando-os, qual apresentou os resultados mais satisfatórios?

- ☐ A busca livre, na qual digitei o que precisava.
- ☐ Utilizando a lista de assuntos definidos pelo *site*.

9.2 Desejaria fazer algum comentário em relação a este estudo ou sobre a pesquisa no *site* da Livraria Cultura?