UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO
MESTRADO EM ADMINISTRAÇÃO - UFRGS / Urcamp

Um estudo sobre a demanda de informações em *sites Web*: o caso de uma unidade de pesquisa de uma empresa de P&D para o agronegócio brasileiro

Ricardo Martins Bernardes

Dissertação apresentada ao Curso de Mestrado em Administração da UFRGS, Escola de Administração, Programa de Pós Graduação em Administração, em convênio com a Urcamp.

Orientador: Professor Dr. Henrique Freitas

Bagé-RS, maio de 2001

Dedico esta dissertação

- aos meus pais,
Luiz Carlos Bernardes (1928-2001) e Abegail Martins Bernardes (1929-1962)

- aos meus filhos,
Alexandre, Michele e Vanessa

- à minha esposa,
Patrizia Ana Bricarello.

Agradecimentos:

À Universidade da Região da Campanha (Urcamp), pelo apoio integral à minha vida. Sem dúvida, minha existência não seria a mesma que é, sem a Urcamp participando de toda a minha vida profissional.

À Empresa Brasileira de Pesquisa Agropecuária - Embrapa - por ser parte indissociável da minha vida pessoal e profissional.

À Embrapa Pecuária Sul, por todo o apoio recebido para a realização deste estudo.

Ao Professor Henrique Freitas, por ter aceitado o compromisso de ser meu orientador.

SUMÁRIO

ANEXO VI : Página do *link* do mecanismo de busca do *site*

GLOSSÁRIO

Backbone - estrutura de comunicações - física e lógica - que possibilita a interligação de redes de comunicações de diferentes organizações/países/continentes.

Banners - publicidade exibida nas páginas *Web*, geralmente utilizando-se de recursos gráficos animados ou não.

Browser - é o programa que "localiza" e exibe uma determinada página eletrônica no computador requisitante. É a parte *cliente* do processo (Netscape, MS-Explorer, HotJava, etc.). O HTTP é a parte *servidor*.

Cookies - registro de preferências dos usuários ao navegar por um *site Web*, que reside no computador do próprio usuário.

Extranet - rede que utiliza a mesma tecnologia da *Internet*, porém com acesso restrito à algumas organizações que se relacionam (fornecedores, principalmente).

Gateway - qualquer computador ou roteador intermediário entre o computador acessado e o computador que acessa.

Hits - unidade de contagem de acessos a páginas eletrônicas. Muito utilizada para se referir a popularidade do *site*.

HTML - *Hyper Text Markup Language* - linguagem utilizada para a elaboração de páginas eletrônicas.

HTTP - *Hyper Text Transfer Protocol* – protocolo que gerencia os acessos às páginas de um *site*. É a parte *servidor* do processo. O *cliente* é representado pelo *browser*.

Hypertexto - é uma metáfora para a apresentação de informações nas quais textos, imagens, sons e ações ficam interligados em uma teia complexa e não linear de associações que permitem ao usuário percorrer assuntos interrelacionados independentemente da ordem em que os tópicos são apresentados. O termo foi criado por Ted Nelson em 1965. Atualmente o termo *hypermidia* tem sido usado em razão dos componentes não textuais do hypertexto, como som, vídeo, imagens e etc.

Intranet – rede que utiliza o mesmo protocolo da *Internet* (TCP/IP), porém com acesso restrito aos membros de uma mesma organização.

ISDN - *Integrated Services Digital Network*. uma rede mundial de comunicações eletrônicas que evolui a partir dos serviços telefônicos existentes. O objetivo da ISDN é a total

digitalização do sistema de comunicações, sem a necessidade de conversão de sinais.

Kbps - *Kilo bits per second* (milhares de bits por segundo). unidade utilizada para medir a taxa de transferência de dados por segundo entre dispositivos e redes. Cada kbps eqüivale a 1.024 bits (128 caracteres).

Links - elementos de ligação que compõem o hypertexto, permitindo a navegação entre diferentes documentos.

Log - registro de transações efetuadas por sistemas de computador. Estes registros podem dizer respeito aos processos internos do sistema, bem como aos processos de interação do sistema com o mundo exterior.

Mbps - *Mega bits per second* (milhões de bits por segundo). unidade utilizada para medir a taxa de transferência de dados por segundo entre dispositivos e redes. Cada mbps eqüivale a 1.024.000 bits (128.000 caracteres).

MS-Windows - sistema operacional da Microsoft para computadores pessoais (PC).

OpenWindows - interface gráfica para sistemas Unix.

Proxies - semelhante aos *gateways*, os *proxies* possuem funções específicas entre o cliente e o servidor. É utilizado normalmente para manter uma memória de todo o conteúdo que circula nas linhas de comunicações a fim de diminuir o tráfego de dados (*cache*).

Robot – um programa que roda de forma automática, sem a intervenção humana. Tipicamente um *robot* é dotado de alguma inteligência artificial, de forma que possa reagir a diferentes situações que ele pode encontrar. Dois tipos de *robots* são agentes e *spiders*.

Spider – ver *robot.*

SQL - *Structured Query Language*. linguagem padrão para operações em sistemas gerenciadores de banco de dados (SGBD).

TCP/IP - acrônimo de *Transport Control Protocol/Internet Program.* um protocolo de software base da *Internet,* desenvolvido pelo Departamento de Defesa dos Estados Unidos para a comunicação de computadores.

URL - *Uniform Resource Locator*. como são chamados os endereços *Web* (http://, ftp://, etc).

Wireless - tecnologias para comunicação sem a utilização de cabos físicos.

X-Windows – interface gráfica para sistemas Unix.

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Foi realizado um estudo de caso exploratório, visando mapear a demanda por informações em um *site Web* mantido por uma organização de P&D para o agronegócio. Os dados utilizados consistiram na seqüência de *clicks* (*clickstream*) realizados por visitantes entre as páginas do *site*, e nas palavras-chave (*keywords*) inseridas no seu mecanismo de busca. Procurou-se abordar a análise considerando as estatísticas gerais de acesso oriundas do domínio ".br", as primeiras ações dos visitantes ao acessar a página principal e as necessidades explícitas por informações, simbolizadas pelas palavras-chaves utilizadas para busca no *site*. A discussão dos resultados foi orientada à descoberta de conhecimentos que pudessem elevar o nível de personalização e customização do *site* estudado, ressaltando-se, entre os quais, o mapeamento das origens das visitas ao *site*, da distribuição das sessões ao longo de diferentes janelas de tempo, das preferências primárias de navegação dos visitantes através dos *links* do *site* e das necessidades explícitas por informação relacionada ao tema do *site*. Considerando o contexto atual, onde cada vez mais as organizações tenderão a buscar e interagir com clientes e parceiros através da *Web*, pode-se concluir que o estudo revelou aspectos úteis relacionados as atividades de transferência de tecnologia e marketing para a organização em estudo.

ABSTRACT

This work reports on an exploratory study carried out on a Web site maintained by an organization of R&D applied to brazilian agribusiness. The analisis covered both the clickstream followed by the visitors and the searches for specific keywords using the site engine. The objective of the study was directed towards the discovery of knowledge that could be used for personalization and customization of the site, focusing on session origins and their distribution along time, primary navigation preferences and on explicit necessities of information related to knowledge organization domain. The results revealed useful aspects which could be applied to technology transfer and marketing activities in the organization, specially considering the current context of viewing the Web as a marketplace where the organizations will have to look for and to interact with customers and partners.

CAPÍTULO 1 : Audiência em *Web sites:* em busca do *feedback* e da personalização

1.1 Justificativa e Tema

Acompanhada por interfaces cada vez mais amigáveis, a rede *Web* tirou a *Internet* do meio técnico-científico para torná-la útil também às pessoas comuns, às empresas, aos governos e organizações com as mais variadas finalidades, impactando rapidamente a economia mundial.

Sua rápida e exponencial expansão se contrapõe a lenta capacidade dos governos em regulá-la. Isto vem contribuindo também, para a tendência de globalização, uma vez que pressiona - através do comércio eletrônico e do intercâmbio de informações - as barreiras econômicas, sociais e políticas que impedem a livre troca de tecnologias, serviços e produtos entre países, homogeneizando as regras de mercado em âmbito global. Por sua vez, é razoável admitir que a globalização representa um dos fatores que acirram a competitividade.

Com a palavra competitividade inserida firmemente no contexto empresarial dos últimos anos, assiste-se à uma atenção crescente das empresas com sua presença na *Web*. O comércio eletrônico, em particular, sinaliza para a mudança do conceito tradicional de "horário de expediente" das empresas. Os espaços virtuais são concebidos para funcionarem independentes da presença de pessoas no ambiente da organização. Desta maneira, nesta nova mídia as empresas nunca fecham, ficando disponíveis e acessíveis para que clientes obtenham informações de seus produtos e serviços, efetuem compras, transações, reclamações e sugestões durante 24 horas por dia e nos sete dias por semana. A presença das empresas na *Web* tornou-se obrigatória, deslocando a competitividade também para aquele espaço.

Assim, a *Web* vem representando, cada vez mais, uma poderosa ferramenta para promover a imagem da organização e estabelecer um canal de comunicação eficiente e um ambiente propício para a satisfação do cliente.

Considerando esta tendência, tornou-se comum a prática das empresas divulgarem em sua publicidade também seus endereços na *Web*. Observa-se que esta tendência abrange todo espectro das organizações, sejam elas privadas ou públicas, filantrópicas, comerciais, industriais, de serviços, de ensino, de pesquisa e etc. Ultimamente a tendência de "possuir endereço na rede" tem alcançado, também, pessoas físicas, configurada através dos "*sites* pessoais"[1].

Para que a ferramenta possa apresentar resultados eficazes em empresas orientadas ao mercado é necessário, entretanto, que ela esteja alinhada com o negócio da organização. Este alinhamento é possível, desde que se efetuem pesquisas junto aos visitantes – usuários e clientes em potencial - a fim de investigar suas preferências e os processos utilizados para a busca e a recuperação de informações no *site*, com o objetivo de orientar sua evolução.

Assim, nos últimos anos os *sites Web* têm experimentado significativas melhorias no sentido de tornarem-se adaptativos, agradáveis e chamativos às pessoas. Também tornou-se obrigatório o uso de mecanismos de captura de informações dos visitantes, como dados cadastrais, sócio-econômicos, preferências e impressões a respeito dos produtos e serviços da empresa. Ambientes sensíveis a escolhas realizadas pelos usuários durante à navegação pelos *sites*, tornam-se cada dia mais sofisticados. Isto inclui *banners* dirigidos e utilização de *cookies*[2], entre outros.

O desenvolvimento de ferramentas e técnicas visando a personalização das relações entre o usuário e o *site* acessado tem sido um dos focos de pesquisas acadêmicas recentes. Estas se dividem em várias linhas visando, entre outros objetivos, proporcionar ao usuário a recuperação de informações relevantes, investigar qual o seu comportamento ao atravessar um *site Web*, quais são suas estratégias de busca de informação e suas preferências frente aos tópicos apresentados nas páginas. Os fundamentos destas pesquisas se baseiam, fortemente, nas linhas de pesquisa da multimídia e da inteligência artificial. Dividida entre o direito à

---

[1] www.registro.br/estatisticas.html, consulta em 04/2000.

[2] http://www.w3.org/Protocols/rfc2109/rfc2109, (HTTP State Management Mechanism) consulta em 04/2000.

privacidade dos usuários e a necessidade de ter *sites Web* competitivos, tem-se como principal interessada nestas pesquisas a área de *marketing* das organizações.

Assim, as premissas gerais que fundamentaram a realização deste estudo foram:

a) sendo a rede *Web* um espaço universal onde habitam todos os tipos de interesses, pode-se admitir que existe demanda por informações de qualquer natureza;

b) ao procurar informações na rede o usuário deixa "rastros" de sua navegação, que são capturados e registrados de forma automática pelos servidores de páginas dos *sites* acessados. Os dados deixados pelos usuários podem incluir qual mecanismo de busca ou *site* e qual a palavra-chave que o levou a determinada página da organização. Uma vez dentro do *site* da organização, o usuário pode seguir *links* ou utilizar-se de novos mecanismos de busca - quase sempre restritos ao âmbito do *site* - para obter a informação desejada. Nos dois casos ele, novamente, deixa registros que podem incluir sua trilha pelo *site* e as palavras-chave sobre as quais ele julga que encontrará a informação desejada. Além disso, a origem do acesso, ou seja, a organização ou localização geográfica do usuário, assim como dados de data e hora de acesso a cada página podem ficar registrados no *site* da organização acessada (W3C, 1999);

c) considerando o contexto atual, onde cada vez mais as organizações tenderão a buscar e interagir com clientes e parceiros através da *Web*, pode-se supor que o estudo daqueles dados pode revelar conhecimentos úteis para a adaptação e customização dos *sites*, visando a personalização das relações com os visitantes (BAMSHAD, 2000) e favorecendo diretamente as atividades de *marketing* das organizações (BÜCHNER *et.al.*,1999).

O estudo, que foi efetuado em uma unidade de negócios de uma organização federal de pesquisa e desenvolvimento para o setor agropecuário localizada em Bagé, no Estado do Rio Grande do Sul, Brasil, teve por objetivo analisar os registros de navegação de usuários através de um *site Web* com a finalidade de extrair conhecimento útil para a organização que o mantém. O centro de pesquisa estudado tem a missão institucional de "Atender com soluções tecnológicas as necessidades dos sistemas produtivos integrados ao agronegócio de bovinos e ovinos na Região Sul em benefício da sociedade" (PLANO DIRETOR DO CENTRO DE PESQUISA DE PECUÁRIA DOS CAMPOS SULBRASILEIROS, 2000).

Os resultados obtidos revelaram conhecimentos úteis para a organização, não somente para a configuração do seu *site Web*, mas como subsídio para suas estratégias de negócios e para sua efetiva orientação ao mercado na *Web*.

Para isto, o capítulo 1 apresenta a justificativa, o tema e os objetivos gerais e específicos deste estudo. Na fundamentação teórica - capítulo 2 - o tema é desenvolvido em uma abordagem *top-down*, ou seja, são apresentadas impressões de pessoas expoentes a respeito do fenômeno "rede *Web*". A seguir são exibidos dados históricos de seu surgimento, algumas projeções sobre seu crescimento e a reação das empresas. Finalmente, é abordado o tema principal, que se refere a análise de registros de transações em servidores de informação (*sites Web*). O capítulo 3 aborda aspectos do método utilizado para a condução da pesquisa, bem como a contextualização do caso. No capítulo 4 são apresentados os resultados, orientando-se a discussão à descoberta de conhecimentos que possa elevar o nível de personalização e customização do *site* estudado No capítulo 5 são apresentadas as conclusões e recomendações produzidas pelo estudo.

1. 2 Objetivos

Abaixo são descritos os objetivos do estudo.

1.2.1 Geral

- Mapear a demanda por informações em um *site Web*, através da análise de registros de acessos (*log de transações*), visando sua configuração e evolução.

1.2.2 Objetivos Específicos

- coletar os registros de navegação de usuários através das páginas de conteúdo (*clickstream*), bem como os registros de palavras-chave (*keywords*) utilizadas no mecanismo de busca do *site Web*;

- aplicar procedimentos quantitativos tradicionais visando elucidar as métricas básicas de acesso ao *site Web*;

- analisar e discutir os registros de navegação de usuários através das páginas do *site*, procurando definir suas preferências de navegação e de conteúdo;

- analisar e discutir as necessidades explícitas de consumo de informações, contidas no registro de palavras-chave utilizadas pelos visitantes no mecanismo de busca do *site*;

- tecer algumas considerações visando orientar futuros estudos de análise de *logs*.

CAPÍTULO 2: Fundamentação teórica

Neste capítulo procurar-se-á apresentar uma visão *top-down* do tema enfocado pela dissertação. Através da indução, tentar-se-á ligar as estratégias das empresas e a importância das redes para o contexto de uma nova economia à necessidade de coletar, armazenar e analisar dados sobre a interação usuário - *site Web*, com o propósito de aplicar o conhecimento advindo daquelas análises como forma de inteligência de mercado.

Procurar-se-á caracterizar a presença na rede *Web* como crucial para a sobrevivência das organizações atuais. Para isto serão expostos alguns conceitos básicos que tornam a existência da rede possível, dados de seu crescimento, suas tendências e suas ligações com *marketing*. Serão também apresentadas impressões recentes de pessoas atuantes nas áreas de administração e informação.

2.1 Rede, Estratégia e Sobrevivência

Nota-se que cada vez mais a presença na rede *Web* é colocada como questão de sobrevivência para as organizações atuais. Esta tendência é corroborada pela visão de teóricos e práticos, especialistas e empresários. Os cenários previstos por diferentes visões partilham um aspecto em comum: a presença na rede.

Especialistas em estratégia, como OHMAE (1998, p.6), consideram que estamos vivendo na *Era da Informação*, e que a capacidade da empresa em atingir os consumidores através da *Internet* é essencial. Ohmae coloca ainda que com o comércio eletrônico tornando-se realidade, a empresa não precisa estar no mercado para atingir seus clientes e consumidores. Adverte, entretanto, que "...para que uma companhia sobreviva a era digital deverá entender a tecnologia, as redes e, acima de tudo, a psicologia de seus principais clientes mundiais..." .

Na mesma lógica, PENZIAS (1998, p.30) coloca o advento das redes como a mudança mais

importante ocorrida nas últimas décadas. Em suas palestras afirma que a utilização das redes evoluirá além das fronteiras corporativas, criando o que chama de “empresa expandida”. Sobre sua importância para as empresas atuais, Penzias diz: “...cada conjunto significativo de dados, cada pessoa e provavelmente quase todas as máquinas estão ligadas em alguma forma de rede. Portanto, as empresas que explorarem essas redes para encontrar pessoas e informações serão certamente as vencedoras...”. Esta percepção o levou a trocar o título de suas palestras de “Administração em um mundo de alta tecnologia” para “Administração em um mundo ligado em rede”.

Para CASTELLS (1999) a tecnologia da informação é uma nova revolução tecnológica, um novo paradigma que possui algumas características tidas como aspectos centrais, que no conjunto formam a base material da sociedade da informação. A primeira característica do novo paradigma é que a informação é sua matéria-prima: são tecnologias para agir sobre a informação, não apenas informação para agir sobre a tecnologia como foi o caso das revoluções tecnológicas anteriores. As outras características são a lógica de redes, a flexibilidade exigida às organizações em razão da lógica de redes, e a crescente convergência de tecnologias específicas para um sistema altamente integrado.

Ainda na mesma linha, TAPSCOTT (1999, p.6) afirma que as empresas e os países que não conseguirem administrar a transição para uma nova economia e uma nova tecnologia estarão em perigo. Especificamente no caso das empresas, Tapscott considera que “...as que não conseguirem se transformar em organizações em rede e forem incapazes de criar comunidades de comércio eletrônico deixarão de ser competitivas e definharão até desaparecer...”. Prevê ainda que no ano de 2005 cerca de um bilhão de pessoas terão acesso a *Internet* . TAPSCOTT e CASTON (1995) já haviam formulado uma escala evolutiva para empresas, onde um dos estágios seria a Empresa Expandida, resultado da computação inter-empresa obscurecendo as linhas divisórias entre organizações e viabilizando novas formas de relacionamentos comerciais. Segundo os autores, “...as redes estão sendo ampliadas... transformando a natureza das interações comerciais e levantando questões de extrema relevância sobre as estratégias de negócios”.

Por último, GATES (1999), um dos ícones empresariais deste final de século, prevê que na próxima década a velocidade com que uma empresa reúne, administra e usa a informação

determinará se será vencedora ou perdedora. Afirma que os "... os vencedores serão aqueles que desenvolverem um "sistema nervoso digital", de tal forma que a informação possa fluir com facilidade através de suas empresas, para um máximo e constante aprendizado".

2.2 *Internet, Web* & convergência tecnológica

Neste tópico serão abordados alguns detalhes, números e aspectos históricos da criação e evolução da *Internet* e da *Web*. Embora o assunto esteja amplamente disseminado pela literatura acadêmica e pelos veículos de comunicação de massa (jornais, revistas, televisão e na própria *Internet*), determinados aspectos, considerados importantes ,serão enfocados.

Sabe-se que a expansão da *Internet* tem se dado de forma bastante acelerada sendo, atualmente, um fenômeno global. A proliferação de provedores de acesso, as possibilidades de acesso via satélite, via estrutura de TV a cabo, as melhorias nas malhas físicas tradicionais de comunicação e o barateamento dos microcomputadores estão contribuindo para o aumento da sua capilaridade e, conseqüentemente, para que um número maior de pessoas se liguem aquela rede. A *Internet* pode ser descrita como uma grande malha de comunicação digital interligada por computadores (*gateways*), que para se comunicarem partilham um protocolo comum chamado *TCP/IP*. Seu nascimento é atribuído à necessidade de aplicações militares, mais tarde expandindo-se para comunidades de ensino e pesquisa americanas. Atualmente experimenta um crescimento acelerado em razão das aplicações comerciais[3].

O surgimento da *Web* (também conhecida como *Word Wide Web, WWW ou W3*) deu-se aproximadamente 20 anos após ao advento da *Internet*. O protocolo base da *Web* foi implementada em 1990 por Tim Berners-Lee e Robert Gaillian, pesquisadores do Laboratório Europeu de Física da Partícula - CERN. A idéia inicial era elaborar uma interface amigável, que possibilitasse a qualquer usuário o acesso a informações de seu interesse utilizando o conceito de *hypertexto*. Para isto desenvolveram os conceitos básicos de *URL*, *HTTP* e *HTML*[4].

A *URL - Uniform Resource Locator* - é o endereço do computador onde está armazenado a

[3] http://www.internet.org e http://www.isoc.org/internet/history/brief.html
[4] http://www.w3.org/History/

página ou recurso. O *HTTP - Hypertext Transfer Protocol* - é o protocolo que gerencia o envio das páginas ao requisitante e *HTML - Hypertext Markup Language* - é a linguagem que define como as páginas devem ser remontadas pelo requisitante. A remontagem é realizada por um *browser* instalado no computador requisitante. O primeiro *browser* gráfico (*Mosaic*) surgiu em 1993, criado do Marc Andreessen, na época programador do *National Center for Supercomputing Applications - NCSA* e, mais tarde, um dos sócios da *Netscape Comunication*, uma das empresas que, juntamente com a *Microsoft*, dominam o mercado daqueles programas.

Resumidamente, a forma tradicional de acesso à *Web* pode ser assim descrita: um computador, com algum sistema operacional, com uma interface gráfica, um *browser* instalado, uma interface de rede ou modem conectados a uma rede que acesse a *Internet*, pode acessar páginas eletrônicas localizadas em qualquer outro computador, também em rede, que tenha um programa servidor de páginas baseado no protocolo *HTTP*. As páginas são transmitidas pelo servidor em forma de comandos (*tags*) em *HTML* e, à medida que chegam ao computador requisitante, vão sendo remontadas pelo *browser* que, para isto, obedece as formas descritas pelos tags enviados pelo servidor. Conexões separadas são abertas para outros elementos como imagens e sons. Comumente esta descrição é representada na forma abaixo (Figura 1):

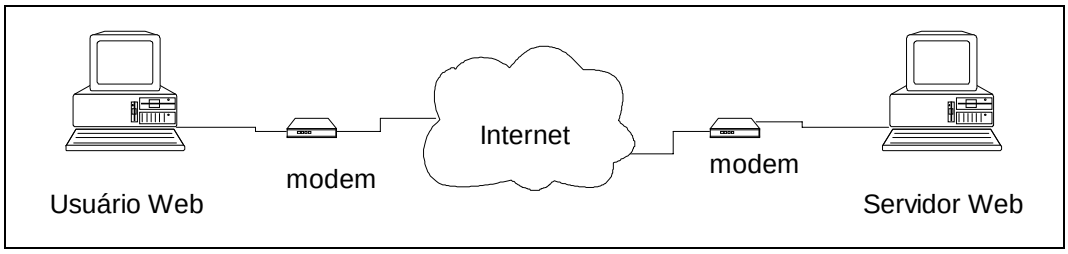

FIGURA 1 - Representação da forma de acesso tradicional na rede *Web*

No entanto, a descrição acima merece algumas considerações: computadores de diferentes arquiteturas de hardware e sistemas operacionais podem acessar a rede e, atualmente, aparelhos de televisão, telefone (*Web Video Phone*), refrigeradores, fornos de microondas e outros aparelhos também podem acessar a *Web*. As formas de acesso estão evoluindo rapidamente, destacando-se as tecnologias *Wireless*, *ISDN*, *ADSL*, *cable modem*, fibra ótica e,

recentemente, viabilização do acesso via rede elétrica. A interface gráfica pode ser o *MS-Windows*, *OpenWindows*, *X-Windows* ou outra. O *browser* pode ser o *Netscape Communicator* ou *MS-Explorer*, ou outro menos conhecido. O que torna possível que equipamentos diferentes, com sistemas operacionais diferentes e conectados através de diferentes meios físicos troquem informações é o protocolo *TCP/IP*, o protocolo padrão de comunicação da rede *Internet*. Isto nos leva a pensar em um novo modelo para representar o processo de acesso a rede *Web* (Figura 2):

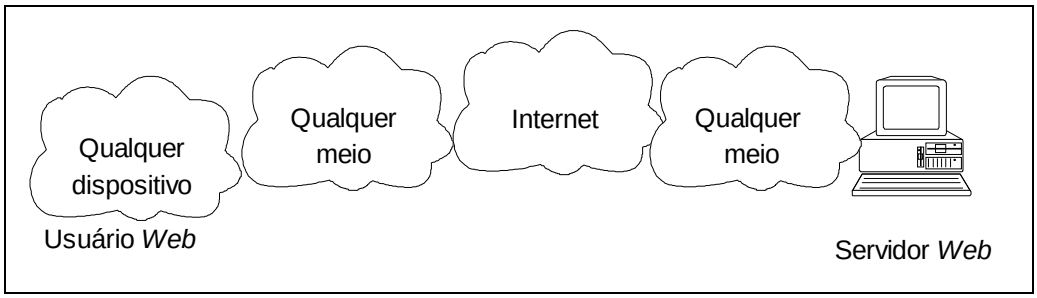

FIGURA 2 - Uma representação mais atual da forma de acesso a rede *Web*: convergência

Para o Grupo de Implantação do Programa Sociedade da Informação no Brasil (SocInfo/MCT) três fenômenos interrelacionados estão na origem das transformações em curso: a convergência da base tecnológica - em decorrência da digitalização da informação, possibilitando sua transformação em conteúdo transmitível; a dinâmica da indústria - que tem proporcionado contínua queda de preços dos microcomputadores, colaborando para popularizá-los - e o crescimento exponencial da Internet.

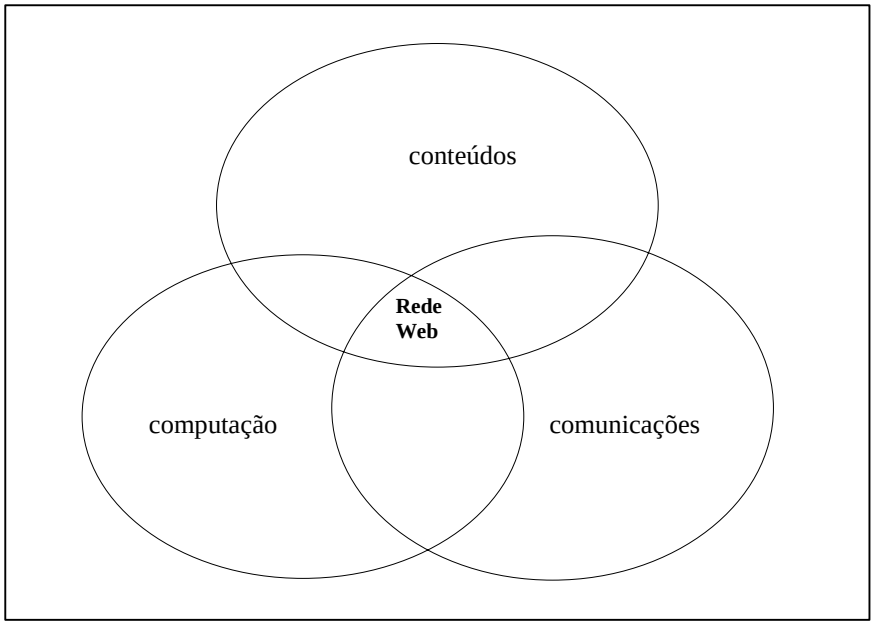

FIGURA 3 - Convergência de Conteúdos, Computação e Comunicações

FONTE: SocInfo 2000, adaptado de http://www.socinfo.org

Apesar da importância da *Internet*, BERNES-LEE (1999) - um dos criadores da *Web* - diferencia a *Internet* da *Web* da seguinte forma:

> "... a Web é um espaço abstrato (imaginário) de informação. Na [Inter]Net, as conexões são cabos entre computadores; na Web, conexões são links de hypertexto. A Web existe por causa dos programas os quais se comunicam entre computadores na [Inter]Net... A Web fez a [Inter]Net útil por que as pessoas estão realmente interessadas em informação e não querem, na verdade, ter que saber sobre computadores e cabos..." .

2.3 O fenômeno da rede *Web:* popularização exponencial

É o crescimento da capilaridade rede *Internet* que possibilita a expansão da rede *Word Wide Web*. Enquanto a primeira estabelece os conceitos de como a informação trafega pela rede - tendo como base o protocolo *TCP/IP* -, a segunda estabelece os padrões de apresentação da informação ao usuário - tendo como base o protocolo *HTTP*. O protocolo base da *Internet* foi criado em 1974 por Vinton Cerf e Bob Kahn. Já o protocolo base da *Web* foi criado em 1990 por Tim Bernes-Lee. Nesta lógica, uma vez que a segunda utiliza-se da estrutura e dos protocolos da primeira para existir, a *Web* cresce porque a *Internet* cresce. Na seqüência, serão apresentados alguns dados com a finalidade de fornecer uma idéia de sua expansão.

Em 1980, a espinha dorsal do que hoje é a *Internet* funcionava a 56 Kbps. Esta velocidade já era de 45Mbps, em 1990, o que representa um crescimento de mais de 800 vezes em 10 anos. Com a implementação da *Internet2*, chegou-se à velocidades superiores à 100 Mbps. O microcomputador, lançado em 1976, espalhou-se pelo mundo de forma muito rápida. Além do preço, o fator utilidade está sendo determinante para sua disseminação em massa.

Com a estrutura de comunicações em expansão e a crescente massificação de microcomputadores pessoais, em meados de 1997 o número de equipamentos ligados na *Internet* dobrava a cada 15 meses e o número de servidores da teia mundial *Word Wide Web* duplicava a cada 14 semanas. Como conseqüência , o tráfego de dados dobrava a cada 100 dias.

Atualmente, o número de pessoas conectadas na *Internet* no mundo continua a crescer aceleradamente. Pesquisas elaboradas por órgãos especializados estimavam em 201 milhões de usuários em todo o mundo, em setembro de 1999. Eram 16 milhões em dezembro de 1995. No Canadá e nos Estados Unidos concentra-se a grande maioria (112,4 milhões), enquanto na América do Sul apenas 5,29 milhões de pessoas estavam conectadas em 1999. Entretanto, previsões apontam 24,3 milhões de usuários da *Internet* na América Latina em 2003[5].

[5] http://www.nua.ie/surveys/how_many_online/index.html, consulta em 13/10/1999.

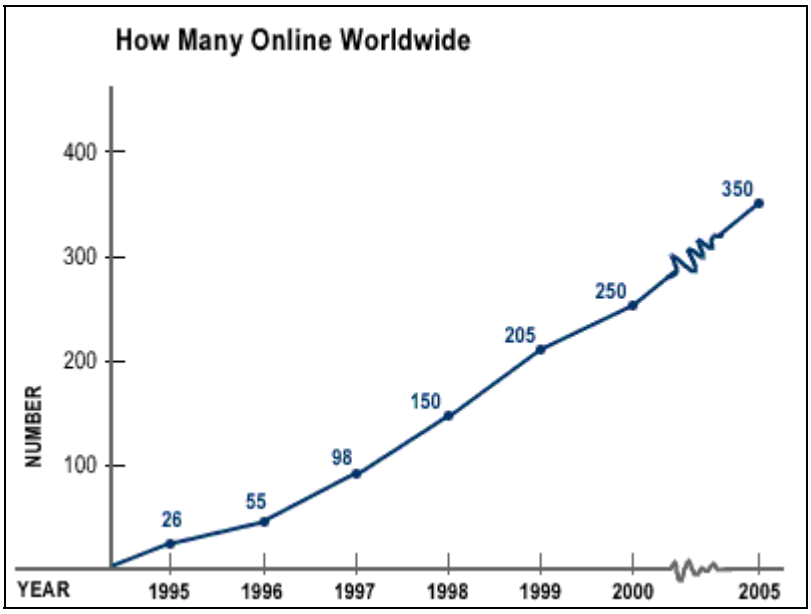

FIGURA 4 - Estimativas de usuários da *Internet* no mundo no período 1995-2005 (em milhões)
FONTE: Nua Surveys, 1999

No Brasil, pesquisa conduzida pelo IBOPE em nove regiões metropolitanas (36,9 milhões de habitantes) em junho de 1999, estimava em 3,3 milhões de usuários. Segundo a mesma pesquisa, entre dezembro de 1998 e junho deste ano, 750 mil pessoas aderiram a *Internet* – um crescimento de 30% - , apontando o Brasil como um dos países que mais usavam a rede no mundo. Apesar da crise cambial, o que elevou o preço dos microcomputadores, os fatores do crescimento estão sendo creditados à expansão da oferta de linhas telefônicas, a tendência de barateamento dos preços das assinaturas de provedores e a notoriedade da rede em função da divulgação nas mídias tradicionais. O aumento do acesso da classe C à *Internet* subiu 33%, enquanto das classes A e B foi de 20%. Em janeiro de 1996 apenas 170,4 mil pessoas tinham acesso à *Internet* no Brasil[6].

Entretanto, os motivos que estão fazendo a rede crescer no momento não são os mesmos que levaram a sua criação. De aplicações militares, no inicio, passando para aplicações acadêmicas de cunho técnico e científico em um determinado momento, atualmente a rede

[6] http://www.ibope.com.br, consulta em 13/10/1999.

tornou-se a vitrine do comércio mundial. O advento da *Web* representou o surgimento de um novo canal para as organizações anunciarem e venderem seus produtos e serviços, difundir sua imagem e ter contato com seus clientes, usuários e consumidores.

Conforme um recente relatório do IDC, a economia em torno da *Internet* movimentará US$ 1 trilhão em 2001, devendo alcançar US$ 2,8 trilhões no final 2003. Para o ano corrente, o mesmo relatório estima movimentação de US$ 507 bilhões[7].

Outro relatório estima que, no mundo todo, o comércio *on-line* entre as empresas e seus consumidores (*business-to-consumer)* deve movimentar US$ 380 bilhões em 2003. Só nos Estados Unidos serão US$ 108 bilhões. Já as transações entre empresas (*business-to-business)* movimentarão US$ 1,331 trilhão no mesmo período. Em 1998 estes números foram de US$ 8 e US$ 43 bilhões, respectivamente. Uma pesquisa da *Forrester Research* estima que serão gastos US$ 30 milhões em publicidade *online* no Brasil em 2004. No mundo todo o total de gastos em publicidade na *Web* será de US$ 33,1 bilhões[8].

É esta força que faz a *Internet* crescer atualmente. A participação de domínios comerciais na *Internet* cresceu de 1,5% em meados de 1993 para 62,6% no início de 1997. Em outubro de 1999 esta participação representava mais de 90% no Brasil. Entre setembro de 1998 e outubro de 1999, o número de domínios .COM.BR cresceu de 107,6 para 115,3 mil. Em janeiro de 1996 eram apenas 479. Até outubro de 1999 a FAPESP, órgão administrador dos domínios da *Internet* no Brasil, contabilizava mais de 130 mil *sites* registrados no País[9].

Um estudo recente da Cyveillance, denominado “The size of Internet” estimou a existência de 2,1 bilhões de páginas únicas com acesso público em novembro de 2000. O estudo indicava que, crescendo sete milhões de páginas ao dia, o número deveria dobrar em meados do ano de 2001[10].

Apesar do crescimento e da importância cada vez maior para o público urbano, um estudo realizado em 1999 pela Research International Brasil, a pedido da Associação Brasileira de

---

[7] http://www.idc.com, consulta em 11/1999.
[8] http://www.nua.ie/surveys/graphs_charts/comparisons/ecommerce_us.html, consulta em out/1999.
[9] http://www.registro.fapesp.br, consulta em out/1999.

Marketing Rural, mostrava que o acesso à informática pelo produtor rural brasileiro ainda é modesto. De acordo com o estudo, 18% dos produtores rurais possuíam computadores e apenas 4% acessavam a Internet. Apesar de tímidos, os resultados mostram o potencial de crescimento para a informática no meio rural[11].

2.4 A resposta das empresas

Naturalmente, À medida que a era da informação e do conhecimento se materializa, a rede torna-se um canal efetivo de realização de negócios (*e-commerce*), com o elemento competitividade deslocando-se para o espaço virtual proporcionado pelas redes. Face a esta realidade, as tecnologias e estudos sobre suas aplicações se sucedem rapidamente.

O emprego intensivo da tecnologia da informação como ferramenta indispensável para as organizações se adaptarem de forma dinâmica ao seu mercado já é considerado um pressuposto básico. Entretanto, estes recursos devem ser usados não somente para suprir as necessidades operacionais e táticas da organização, mas também, serem efetivamente utilizados pelo nível estratégico.

Quanto a isto, FREITAS (1993) tece várias considerações sobre o que uma empresa deve fazer para que seja competitiva na era da informação e do conhecimento. O autor considera que a empresa deve adaptar-se ao cliente e antecipar suas expectativas, desenvolver sua capacidade de evolução para assegurar sua perenidade, desenvolver sua capacidade de escuta do exterior e de seu próprio futuro para melhor adaptar-se, gerindo a informática como um meio integrado na estratégia.

Nesta linha, GONÇALVES e FILHO (1995) previram que, com a redução dos custos dos equipamentos, a expansão das redes e desenvolvimento do *software* e *hardware*, as possibilidades de se aplicar a tecnologia da informação no *marketing* aumentarão a cada dia. Neste contexto, ganharão as empresas inovadoras que souberem utilizar a TI para melhorar o relacionamento com os clientes, na busca de vantagem competitiva. Segundo os autores, as

---

[10] http://www.cyveillance.com/newsroom/pressr/000710.asp, consulta em nov/2000.
[11] Technonews - On Line (http://www.technovet.com.br/tnews, ano IV - n.46 - Out. 1999.

inovações na TI estão voltadas para os sistemas de apoio ao cliente (SAC). Esta inovação pode dar-se através de pesquisas de mercado e serviços de informações para inteligência competitiva, entre outras.

NOVAK e HOFFMAN (1995) apresentaram fundamentos conceituais para o *marketing* em ambientes *hypermidia*, visando introduzir a área de *marketing* na mídia então emergente (*Web*)[12] e propor um modelo estrutural de comportamento de consumidor em ambientes de comércio *hypermidia*. No mesmo trabalho os autores examinaram um conjunto de proposições de pesquisa resultantes do modelo proposto e delinearam questões chaves para a pesquisa, resultantes da necessidade de investigar a área emergente.

Em um trabalho posterior HOFFMAN e NOVAK (1996) compilaram os novos conceitos advindos da utilização da *Web* como ferramenta de *marketing* e propuseram uma metodologia e uma terminologia padronizada para medidas de publicidade na *Web*. As métricas sugeridas pelos autores se constituem em definições formais de métricas para tempo de exposição e métricas para a interatividade. Incluem também definições adicionais de medição de audiência de publicidade na *Web*, dentre os quais estão os padrões primários de navegação através de um *site Web*, padrões de navegação entre *sites*, características demográficas, psicográficas e comportamentais dos visitantes em um *site* e de páginas específicas dentro do *site*, medidas de atitude e cognição frente ao conteúdo e medidas de lealdade do visitante e re-visitas ao *site Web*. Recomendações preliminares para a pesquisa sobre publicidade na *Web* também foram sugeridas.

Com sua rápida expansão e com as possibilidades de exploração comercial, nos últimos cinco anos, a *Web* transformou-se em uma forma básica de relacionamento entre as empresas e seu mercado. Analisando pesquisas relacionadas, observa-se que o assunto tem se tornado objeto de estudo de pesquisadores e profissionais que seguem os mais diversos focos de investigação.

SALAM *et.al.* (1996) realizaram pesquisa em 45 *sites* de empresas americanas incluídas no *Fortune 1000*, utilizando quatorze critérios - ou sugestões - elaborados por RESNIK e

[12] Na verdade, os autores usam o termo "CME" – *Computer Mediated Environment.*

STERN (1977)[13] , citado pelos autores. Os critérios avaliados foram existência de preço ou valor, qualidade, performance, componentes ou conteúdo, disponibilidade, oferta especial, gosto (estética), forma, garantia, segurança, nutrição, pesquisa independente, pesquisa apoiada pela empresa e novas idéias. Os resultados evidenciados pela pesquisa mostraram que 85,71% das páginas eram informativas, contendo no mínimo um dos critérios avaliados. Somente 64,29% das páginas continham no mínimo dois dos critérios e 57,14% continham no mínimo três critérios. A pesquisa mostrou ainda que, para os usuários, o critério mais freqüentemente buscado era a performance, seguida de perto por preço e qualidade. Entretanto, apenas 48.9% das páginas examinadas continham estes três critérios.

Em pesquisa conduzida entre 101, das 500 maiores e melhores de 1997 da Revista Exame, SOARES e HOPPEN (1998) concluíram que os *sites* estudados refletiam uma tendência predominante com a imagem corporativa, o que revelava estarem em estágios primários em termos de evolução do uso da *Web*. Esta conclusão está apoiada no modelo evolucionário para as grandes empresas do ramo tradicional que iniciam um *site Web*, proposto por QUELCH e KLEIN (1996)[14], citado pelos autores, e que consiste em cinco estágios: 1) imagem da empresa e informações sobre produtos, 2) coleta de informações e pesquisa de mercado, 3) suporte ao consumidor e serviços, 4) suporte interno à empresa e serviços e 5) transações via rede.

Outros dados da mesma pesquisa merecem ser observados: apenas 19% das empresas possuíam um mecanismo para receber *feedbacks* de clientes em seus *sites*. Ou seja, três entre cada cinco empresas da amostra, não estavam interessadas na opinião do visitante *Web*, ou não viam na *Web* um meio potencial para transformar dados daqueles que acessam seus *sites* em oportunidades de negócio; 55% das empresas não possuíam qualquer estrutura para recolhimento de dados do visitante do *site;* 77% não explicitaram seus clientes alvos; 22% da comunicação era destinada à clientes institucionais e 18% à clientes individuais; dos que tinham algum mecanismo para coleta de dados de visitantes, 98% pediam dados de identificação, 35% dados sócio-econômicos e 15% dados de hábitos ou preferências de

---

[13] RESNIK, A. e STERN, B. Na analysis of information content in television adversing. Journal of Marketing, January 1977, pp. 50-53.
[14] QUELCH, J. e KLEIN, L. The internet and international marketing. Boston: Sloan Magagement Review, Spring 1996.

consumo. Ainda, 89% não ofereciam nenhum incentivo para os visitantes deixarem seus dados.

Outra pesquisa, efetuada por SILVA (1997), na qual foi avaliada a presença empresarial das empresas brasileiras na *Web*, constatou que a grande maioria das empresas incentivava o *feedback*, respondendo sempre as solicitações de seus clientes. Entretanto, poucas estavam se utilizando do *marketing* de relacionamento. Apenas 17% utilizavam *database*, 8% utilizavam estudos de *prospects* e 29% efetuavam pesquisas com clientes. Os valores foram considerados inexpressivos pelo autor.

Na época (1997), a pesquisa de Silva constatou que a *Web* no Brasil estava na etapa inicial do comércio eletrônico, com os *sites* fornecendo prioritariamente informações sobre a empresa (85%) e seus produtos (92%). Apenas 21% das empresas faziam suporte aos clientes e 13% efetuavam suporte aos produtos. As promoções eram efetuadas em apenas 19% das empresas pesquisadas. Confirmando a pesquisa de SOARES e HOPPEN (1997) o autor constatou também que a maioria das empresas mantinha sua presença na *Internet* como forma de solidificar sua imagem institucional e também como fonte de negócios. Cerca de 71% das organizações já realizavam estudos de freqüência dos usuários as suas páginas. Para isto, utilizavam relatórios de freqüência e contagem de *hits*.

Certamente a situação relatada nestes estudos apresenta-se distinta atualmente, mais de três anos após as pesquisas. No entanto, já estarão as empresas, efetivamente, utilizando dados capturados em seus *sites Web* como ferramenta de inteligência de mercado ? De fato, se conseguirmos entender as intenções das pessoas através das suas escolhas ao navegar por um *site Web*, poderemos utilizar esta informação como vantagem competitiva.

MONTGOMERY e WEINBERG (1998) consideram que a análise de consumidores e não consumidores é, possivelmente, a mais valiosa e a mais negligenciada área da inteligência estratégica. Análises de consumidores e não consumidores podem revelar tecnologias emergentes, vantagens e desvantagens competitivas e novas idéias e produtos. Relataram um estudo sobre inovação tecnológica, no qual 74% das 137 inovações estudadas foram originadas pelos consumidores, três vezes mais que os departamentos de pesquisa das empresas. Segundo os autores, seis funções de transformação pelas quais os dados podem se

tornar informação – transmissão, acumulação, agregação, análise, reconhecimento de padrões e mistura - variam consideravelmente em complexidade

2.5 Usuários, suas Trilhas & Padrões de Comportamento

Na retaguarda dos ambientes de *hypermidia*, muitos estudos têm sido realizados no sentido de investigar como o usuário se comporta frente a ambientes mediados por computadores e quais as estratégias para a busca e recuperação de informações de seu interesse. Alguns resultados destas pesquisas servem para realimentar a interface das aplicações, tornando-as mais eficientes no processo de comunicação usuário X sistema de informações. Ultimamente, com o advento do comércio eletrônico, o foco dos estudos está sendo deslocado para a interação usuário X *Web*, tendo a área de *marketing* como uma das principais interessadas.

Em um trabalho anterior a popularização da rede *Web*, FREITAS (1993) propôs um modelo para a avaliação de um sistema de apoio a decisão (SAD). O modelo baseava-se em um método implícito e automático de coleta e armazenamento de todas as ações de um usuário frente a um sistema teleinformatizado de *marketing*. Os resultados do estudo conduzido pelo autor permitiram a elaboração de uma tipologia de usuários finais para o sistema utilizado. Para seu estudo, ele utilizou uma aplicação customizada, que permitia o controle total das ações dos usuários na interação com o sistema. Este não é o caso dos sistemas acessados via *Web*, na qual, muitas ações dos usuários não podem ser capturadas. Além disso, a estrutura de navegação de uma aplicação customizada é diferente da estrutura possível em páginas de *hypertexto*. Enquanto na primeira é possível elaborar uma estrutura de menus pré-determinada e finita, a modelo de uma “árvore”, na segunda a estrutura assemelha-se a uma “malha”, que possibilita ao usuário ampla escolha dos caminhos a serem seguidos.

Um trabalho semelhante ao de FREITAS (1993) foi efetuado por SAKAMOTO (1998). O autor elaborou uma biblioteca (*library*) que, inserida em um *browser*, permitia total controle das ações do usuário frente a uma sessão de navegação pela *Web*. Entre as medidas possíveis com a ferramenta, estão a rota tomada para acessar uma página (via *hyperlink*, *bookmark*, ou entrada direta da URL) , o tipo de ação em uma página (impressão ou registro no *bookmark*), o tempo gasto na leitura de uma página e o tempo que uma página leva para ser mostrada em um terminal. O estudo de Sakamoto é útil em diversos aspectos: para provedores de conteúdo

e publicitários provê meios para medidas de audiência e auditagem. *Webmasters* podem entender melhor as tendências de uso como o número de páginas vistas, bem como obter elementos para a estratégia futura do *site*. Para aos usuários fornece elementos para implementar serviços personalizados para recuperação de informações, recomendações e customização.

Uma das diferenças entre os dois trabalhos é que no de Freitas os dados capturados a partir das ações dos usuários eram armazenados no servidor e, no de Sakamoto, os dados ficavam no cliente - muito embora nada impeça de serem enviados a um servidor remoto para arquivamento. Isto, no entanto, teria sérias restrições em relação a privacidade, sendo aconselhável seu uso apenas em situações controladas e com o consentimento dos usuários.

Uma semelhança, é que as duas aplicações geravam registros de transações entre usuário X sistema (*logs*) para posterior análise e utilização.

Neste aspecto específico, ABDULLA *et.al.* (1997) examinaram os *logs* de transações de um servidor de *proxy* visando verificar similaridades entre acessos realizados por instituições educacionais versus acessos da indústria, governo e provedores de acesso comercial. Os autores não encontraram diferenças entre os acessos, considerando nove fatores comuns que eles denominaram *invariants*. Estes fatores foram estabelecidos por ARLITT e WILLIAMSON (1996)[15], citado pelos autores. A não-diferença encontrada nas medidas observadas é importante, uma vez que contribui para a generalização e a obtenção de modelos estatísticos que podem ser utilizados para simulações e estudos de modelagem.

Uma pesquisa mais voltada ao comportamento de usuários foi realizada por JANSEN *et.al.* (1998) usando os *logs* do mecanismo de busca *Excite*. Foi analisado um *subset* de 51.453 consultas efetuadas por 18.113 usuários escolhidas randomicamente entre as consultas efetuadas em 10/03/1997. Os dados revelados pela pesquisa mostraram uma média de 2,8 consultas por usuário, o que denota ações de refinamento nas buscas. Na média, as consultas possuíam 2,21 termos. Mais de 80% das consultas possuíam de um a três termos, sendo que

---

[15] ARLITT, M.F. e WILLIAMSON, C.L. Web server workload characterization: the search for invariants. Proc. SIGMETRICS, Philadelphia, PA, April 1996. ACM, 160-169.

31,46% utilizavam dois termos e 30,81% , apenas um termo. Em relação a construção das consultas, a pesquisa revelou que o uso de operadores booleanos foi baixo. Apenas 9,32% usaram o operador AND. O uso do operador OR foi de 0,26%, o operador AND NOT foi usado 0,23% e os parênteses foram usados em 0,53% das consultas. Mesmo assim, o percentual de incorreções na montagem das consultas com os operadores AND, OR, AND NOT e parênteses foi de 26,3%, 34,85%, 65,83% e 32,23% respectivamente.

SPINK *et.al.* (1998) observaram que usuários com um problema para resolver (*problem-at-hand*) tendem, ao longo do tempo, a procurar no mesmo ou possivelmente em diferentes sistemas interativos (bibliotecas digitais, sistemas de recuperação de informações, serviços na *Web*) por respostas para o mesmo problema de informação, ou a ele relacionado. Segundo os autores, este processo é chamado de *sucessive search phenomenon*. Observaram também que a grande maioria dos usuários tende a empregar estratégias simples de busca. Na mesma pesquisa, os autores relataram também que apenas 5,24% das consultas continham operadores booleanos, tidos como chaves para o refinamento de buscas. Neste levantamento, o número médio de termos informados por usuário/consulta foi de 3,34. Salienta-se que estes dados foram fornecidos pelos usuários, não envolvendo nenhuma técnica de coleta automática.

Não somente descrevendo os dados obtidos através da análise de *logs*, mas agora aplicando a informação obtida, PERKOWITZ e ETZIONI (1997) apresentaram um estudo inovador, o *sites* adaptativos: um *site* capaz de melhorar a si mesmo através da análise de padrões de acesso dos usuários. Em linha semelhante, JOHN e PANAGIOTIS (1998) propuseram um algoritmo para rearranjar a estrutura de um *site Web* a partir da popularidade das diferentes páginas (*Relative Page Popularity – RPP).* Este algoritmo baseia-se, entre outras medidas, em estatísticas armazenadas nos *logs* de transações coletados e armazenados automaticamente pelo servidor de páginas do *site*. Através de um estudo de caso os autores relataram que o algoritmo contribuiu para o aumento de número de acessos ao *site* em que foi aplicado.

O surgimento da rede *Web* e o seu crescimento em função do potencial comercial, leva à convergência de tecnologias computacionais com aplicações cada vez mais estratégicas para as empresas. A evolução e o barateamento do hardware têm proporcionado às empresas o armazenamento de grandes bases de dados. Por sua vez, técnicas que integram estatísticas tradicionais com inteligência artificial (mineração de dados), aliadas a ferramentas de banco

de dados, possibilitam a extração de conhecimento potencialmente útil daquelas bases.

Pressionados pelas necessidades dos especialistas da área de *marketing* das empresas, o tema torna-se objeto de estudos de especialistas em mineração de dados e em banco de dados, e técnicas e ferramentas sofisticadas tem sido desenvolvidas com propósito de extrair conhecimentos para inteligência de mercado a partir dos *logs* de acesso aos *sites Web* das organizações. Todas estas técnicas e ferramentas de software estão enquadradas dentro de uma “nova disciplina” denominada *Web Mining*.

ZAÏANE (1998a) define o termo *Web Mining* como sendo a extração de padrões interessantes, potencialmente uteis e de informação implícita de artefatos ou atividades relacionadas com a *World Wide Web*. O autor divide o termo ainda em três domínios de descoberta de conhecimentos: *Web Content Mining* (mineração de conteúdo), *Web Structure Mining* (mineração de estrutura) e *Web Usage Mining* (mineração de uso), conforme esquema abaixo (Figura 5):

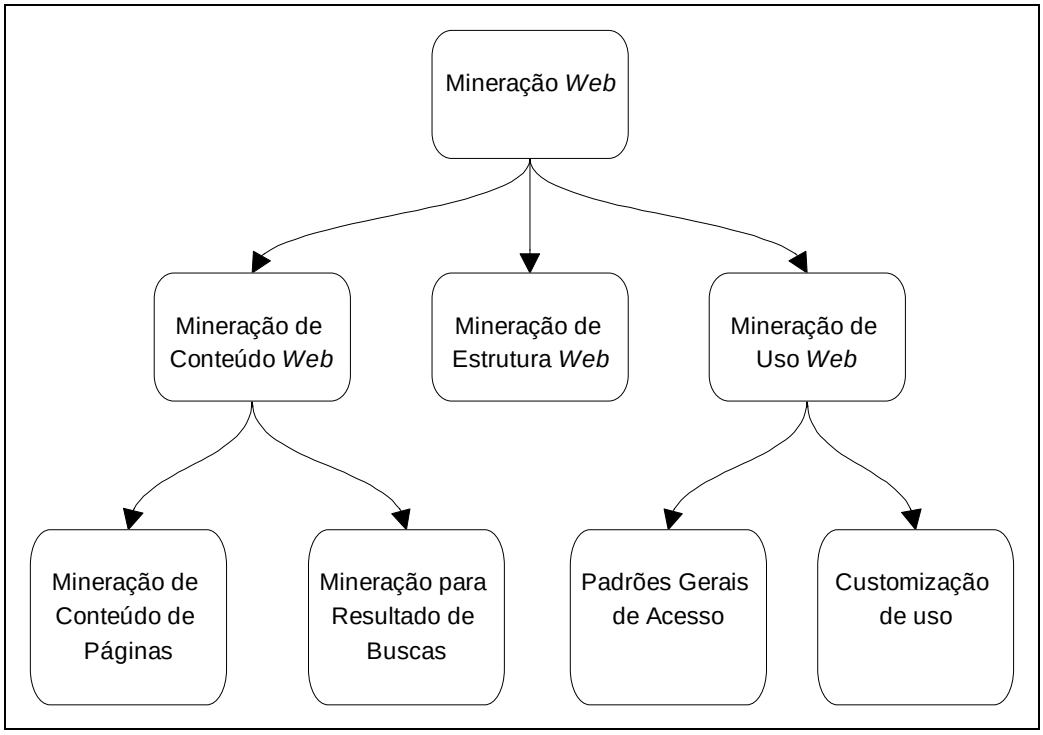

FIGURA 5 - Linhas de investigação em Mineração *Web* (*Web Mining)*

FONTE - Zaïane, 1998a.

A *mineração de conteúdo Web* é o processo de extrair conhecimento do conteúdo de documentos HTML e suas descrições. *Mineração de estrutura Web* é o processo de inferir conhecimento a partir da organização da *Web* e *links* entre referentes e referenciados. Por último, *mineração de uso da Web* - também conhecida como *Web Log Mining* - é o processo de extrair padrões interessantes a partir de *logs* de acesso a *Web*.

Esta última se constitui o foco do presente estudo sendo necessário subdividi-la, ainda, em duas principais tendências: rastreamento de padrões gerais de acesso e rastreamento para customização de uso. Enquanto a primeira visa analisar *logs Web* para entender padrões e tendências gerais de acessos, e com isto melhor estruturar ou agrupar conteúdos ou recursos, a segunda visa customizar *sites Web* para indivíduos, sendo voltada para a personalização de *sites* (*Web Personalization*).

Os trabalhos relacionados foram desenvolvidos em laboratórios de universidades em tempos recentes. Dentre as mais expoentes está a Universidade de Minnesota, Universidade Simon Fraser, Universidade de Alberta e Universidade de Ulster. A seguir será apresentado um pequeno histórico sobre o assunto, bem como relacionados alguns destes trabalhos.

COOLEY *et.al.* (1997a) apresenta o WEBMINER, um software que através da utilização de bancos de dados relacionais e uma linguagem semelhante a SQL *(Structured Query Language*) possibilita ao analista de *logs* a formulação de consultas livres com a fixação de suporte e confiança arbitrários. Técnicas detalhadas e sofisticadas de preparação de dados para mineração de padrões de uso *Web* são descritas em COOLEY *et.al.* (1997b) e COOLEY *et.al.* (1998). Neste último os autores comparam três abordagens distintas de identificação de transações, uma fase em que se agrupam os acessos dando maior significado para a sessão de um visitante, tonando mais eficiente a mineração de padrões.

ZAÏANE *et.al.* (1998b) descreve um protótipo de ferramenta para análise de *logs*. O WebLogMiner utiliza as tecnologias de *data warehouse*, OLAP (*On-line analytical processing*) e de mineração de dados para descobrir padrões de acessos nos *logs* de um *site Web*.

SPILIOPOULOU e FAULSTICH (1998) e SPILIOPOULOU *et.al.* (1999) consideram que a análise do comportamento de usuários possui dois aspectos: um relativo aos interesses do usuário e a informação que eles acessam, e outro relativo ao caminho ou maneira de como a informação é acessada. O primeiro aspecto consiste em estabelecer perfis de usuário, não sendo peculiar a *Web*. Já o segundo possui seu foco nas técnicas de analisar *logs* de servidores *Web*. Para este último, os autores propuseram um algoritmo que foi implementado através de um software (WUM - *Web Utilization Miner*), o qual trilha os caminhos dos usuários através de um *site* utilizando pré-processamento dos *logs*, técnicas sofisticadas de heurística e uma linguagem adicional para livre consulta semelhante a SQL. Entretanto, os autores argumentam que o usuário *Web* é caracterizado pelos seus interesses e pelo seu comportamento navegacional sendo portanto, os dois aspectos, complementares.

Apoiados em uma tendência conhecida como *customização em massa*, BAMSHAD *et.al.* (1999) propuseram uma arquitetura geral para um sistema automático de personalização *Web* baseado em regras de associação e derivação de *clusters* de URL, de transações e de trilhas.

Conferindo um enfoque mais aplicado ao assunto, BÜCHNER *et.al.* (1999) apresentam o MiDAS, minerador de *logs* baseado na arquitetura MIMIC (Mining the Internet for Marketing IntelligenCe), demonstrando como inquirir e aplicar conhecimentos derivados da análise de *logs* para atrair e reter clientes, realizar *cross-sales* e prever a "saída" de clientes.

A literatura apresentada neste capítulo, procurou evidenciar que o surgimento, o crescimento e a popularização da rede *Web*, bem como a sua re-configuração como um canal de promoção, comercialização e distribuição tem levado organizações a marcar presença no espaço virtual, na forma de *sites Web*, questão colocada como estratégica pelo paradigma empresarial contemporâneo.

À medida que aspectos básicos sobre a implementação destes *sites* vão sendo rapidamente automatizados e dinamizados, a fronteira das investigações se desloca em direção a um maior conhecimento da audiência daqueles espaços, empregando a aprendizagem adquirida para vários aspectos, tais como:

a) proporcionar mais foco ao conteúdo;
b) agrupar melhor os conteúdos e recursos;
c) levantar subsídios para profissionais que trabalham com desenho e edição de *sites*;
d) planejar o *site* (estrutura física de pastas, aplicações CGI necessárias, bases de dados, balanço de carga, etc)
e) personalizar *sites*
f) alinhar conteúdo publicado com a estratégia organizacional

Não obstante, as informações obtidas, uma vez conformando com as demandas ambientais, podem se traduzir como forte subsídio para a elaboração das estratégias da organização. Entretanto, a grande diversidade de estruturas e de ferramentas empregadas - além dos aspectos individuais dos desenvolvedores - nos *sites Web* sugere, também, formas diversificadas de abordagens para a realização de investigações, favorecendo e valorizando, de forma especial, os estudos de caso de natureza exploratória e descritiva.

3 MÉTODO DE PESQUISA

Através da revisão da literatura, pode-se inferir que o tema abordado é emergente, carecendo ainda de métodos, técnicas e ferramentas apropriadas para a sua abordagem. A seguir serão descritos os pressupostos metodológicos e as técnicas empregadas para a condução desta pesquisa.

3.1 Nível e Delineamento de Pesquisa

O nível desta pesquisa se enquadra na descrição de GIL (1996, p.45) como exploratória. Conforme o autor, o objetivo deste tipo de estudo é proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou de maneira a possibilitar a construção de hipóteses. Tem como objetivo principal o aprimoramento de idéias ou a descoberta de intuições.

Dadas as suas características, estas pesquisas contribuem para a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores, sendo desenvolvidas para proporcionar uma visão geral, de tipo aproximativo, acerca de determinado fato (GIL, 1999, p.43).

Este estudo, em especial, descreve algumas características de uma população, tentando estabelecer relações entre variáveis utilizadas e informações obtidas. Esta caraterística pode, também, aproximá-lo do nível descritivo (GIL, 1999, p.43).

Uma vez que, segundo GIL (1999, p.72, 1996, p.58) a investigação profunda e exaustiva de um ou de poucos objetos, de maneira a permitir o seu conhecimento amplo e detalhado, é uma das características do estudo de caso, esta foi a estratégia adotada para conduzir a pesquisa. O estudo de caso pode servir a pesquisas com diferentes propósitos, tais como para a exploração de situações da vida real cujos limites não estão claramente definidos, para a descrição de

situações do contexto em que está sendo feita determinada investigação e para a explicação de variáveis causais de determinado fenômeno em situações muito complexas que não possibilitam a utilização de levantamentos e experimentos. Por estas características, o estudo de caso pode ser utilizado em pesquisas exploratórias, descritivas e explicativas.

Ressalta-se, entretanto, alguns preconceitos existentes quanto à estratégia escolhida, tais como a falta de rigor metodológico, a dificuldade de generalização e o tempo de execução de pesquisa muito longo (YIN, 1994, p.9). Todavia, segundo GIL (1996, p.59), o estudo de caso apresenta algumas vantagens, tais como o estímulo a novas descobertas, a ênfase na totalidade e a simplicidade dos procedimentos.

A unidade de análise foi o *site Web* de uma unidade descentralizada, de uma empresa de P&D que atua em nível nacional. A coleta de dados deu-se durante 16 meses, o que caracteriza o estudo como longitudinal (YIN, 1993, p.47). No período de coleta procurou-se seguir os três princípios de coleta de dados apontados por YIN (1994, p.91), os quais são;

a) usar múltiplas fontes de evidência (2)
b) criar uma base de dados do caso e,
c) manter uma cadeia de evidência

Pelo fato de ser um estudo simples, procurou-se a convergência das fontes de evidência utilizadas. Entretanto, ao abordar diferentes aspectos dos acessos ao *site*, o estudo aproxima-se do que YIN (1993, p.103) denomina de “múltiplos sub-estudos”. Já o relatório seguirá a estrutura linear-analítica, descrita em YIN (1994, p.138).

3.2 O Caso - contextualização

A Empresa Brasileira de Pesquisa Agropecuária – Embrapa -, ambiente deste estudo, é uma empresa estatal, fundada em 1973, que vem passando por rápida transformação em sua maneira de atuar, especialmente a partir do início da década de 1990. Sua estrutura consiste em 37 centros de pesquisas e 24 serviços especiais espalhados pelo território nacional que, ao longo dos últimos dez anos vem passando por mudanças contínuas visando sua adequação às reformas do Estado.

Uma das mudanças mais significativas foi a dimensão conceitual assumida de que “a pesquisa começa e termina na sociedade”, em contraposição ao paradigma anterior que pregava que “a pesquisa começa e termina no produtor [rural]”. Documento estratégico orientador da atuação da empresa para a década de 1990 justificava a mudança de enfoque afirmando que “...diferentes segmentos sociais têm demandas diferentes para serem atendidas por diferentes conjuntos de informações técnico-científicas geradas pela Embrapa e ofertadas sobre diferentes formas...” (FLORES, 1991). O mesmo documento colocava como diretrizes a ampliação e fortalecimento das atividades de difusão de tecnologia e a integração com o setor privado, entre outras, pondo como prioridade “...formular e implantar políticas e diretrizes gerais para as áreas estratégicas tais como cooperação nacional e internacional, marketing e comercialização, comunicação social, PeD e captação de recursos...”.

Dentre algumas mudanças implantadas desde o início desta década rumo à orientação ao mercado estão a implantação do SEP (Sistema Embrapa de Planejamento) - que alterou profundamente sua maneira de levantar demandas ambientais para delinear suas atividades de P&D -, uma Política de Comunicação - para delinear o relacionamento com o ecossistema em que atua -, uma Política de Negócios Tecnológicos - para consolidar a nova orientação voltada ao mercado -, a implantação do *SIGER* (Sistema Corporativo de informações Gerenciais) - para gerenciar seus projetos de P&D -, mudanças estruturais em sua cúpula administrativa, adoção de estrutura funcional por projeto, gerência por processos, redução de vagas, redução de custos, redução de níveis hierárquicos e etc.

A introdução do programa de qualidade na empresa remonta ao ano de 1992. Em 1994, publicação interna da empresa, baseada em diversos pressupostos ambientais que pressionavam as instituições de C&T a mudarem a forma de se relacionar com o seu ecossistema, assinalava que “A mudança básica consiste em estabelecer como alvo principal de sua missão e de seus objetivos, os clientes, beneficiários e usuários...” (POPINIGIS et al., 1994). Este pressuposto traz em seu bojo a mudança da orientação ao produto para uma orientação ao mercado, refletindo políticas estratégicas baseadas na tendência de diminuição do papel do Estado em determinadas atividades e a necessidade crescente de geração de recursos próprios.

FLORES e SILVA (1992) observaram que as atividades de *marketing* eram incipientes em nossas instituições de pesquisa, creditando isto a forte orientação ao produto e não ao mercado existente na atividade. Salientaram que era necessário romper com tal paradigma.

KOTLER (1996) considerou que existem cinco conceitos diferentes sob os quais as organizações conduzem suas atividades no mercado: produção, produto, vendas, marketing e marketing societal. Conforme o autor, o conceito de produto assume que os consumidores favorecerão aqueles produtos que oferecem mais qualidade, desempenho ou características inovadoras. Kotler advertiu que esta orientação pode levar os administradores a ter um "caso amoroso" com seus produtos deixando de observar que o mercado pode estar menos preocupado com a qualidade oferecida. Já o conceito de *marketing* (orientação ao mercado) assume que a chave para atingir as metas organizacionais consiste em determinar as necessidades e desejos dos mercados-alvo e oferecer as satisfações desejadas de forma mais eficaz e eficiente do que os concorrentes.

A Embrapa orientou suas atividades baseada no conceito de produto por quase 20 anos, tendo assumido explicitamente a orientação ao mercado a partir de meados da década de 1990. Além da mudança de orientação, veio a clara definição de que a Empresa deveria atuar, como agente do sistema técnico-científico nacional, em pesquisa e desenvolvimento (P&D). Neste aspecto, ZAWISLACK (1996) caracteriza o objetivo da pesquisa tecnológica dos agentes desta classe como "...à procura de aplicação e soluções a problemas econômicos reais, ou seja, diz respeito ao momento da inovação...".

A Política de Negócios Tecnológicos (1998), em particular, consolida direções já expressas em documentos estratégicos da Embrapa, as quais apoiam-se fundamentalmente em *marketing*. Este direcionamento já havia ficado claro no documento Estratégia Gerencial da Embrapa - Gestão 95/98 o qual define como política global de administração que:

> "A Embrapa adota o Marketing, na sua acepção mais ampla de "filosofia de relacionamento com o macro-ambiente"...Isto eqüivale dizer que Marketing não deve ser apenas preocupação ou domínio de um departamento da sede ou de um setor numa unidade descentralizada...é uma atitude nova, um compromisso de todos os empregados da Embrapa para com a sociedade brasileira".

Para amparar esta orientação o documento estabelece três políticas visando atender os quatro compostos de *marketing* (4P'S). Estas políticas eram a Política de Pesquisa e Desenvolvimento (Produto), a Política de Vendas ou Distribuição (Preços e Pontos de Venda) e uma Política de Comunicação (Promoção). O mesmo documento advertia que

> "... a adoção desta nova postura não pode ser retardada por visões segundo as quais marketing não se aplicaria à Embrapa por ser uma empresa que gera, em grande monta, tecnologias de cunho social que não teriam preços e não seriam vendáveis". Para fortalecer esta nova orientação, o documento estabelecia que "... tudo o que se gerar na Embrapa será objeto de rigoroso escrutínio segundo a visão de Marketing ...".

A orientação ao mercado se apóia em quatro pilares básicos (KOTLER, 1996). Estes pilares são mercado-alvo, necessidades dos consumidores [clientes e usuários], marketing coordenado e rentabilidade. A identificação de segmentos de mercado e seleção de mercados-alvo é, portanto, uma etapa obrigatória para a adoção de uma orientação ao mercado. Segundo o autor, as empresas atualmente estão abandonando a prática de mercados de massa por não valerem a pena, pois estes estão sendo pulverizados e sendo transformado em micro-mercados com compradores diferentes, à procura de produtos diferentes em canais de distribuição diferentes. A segmentação de mercado é seguida da escolha do mercado-alvo e o posterior posicionamento, quando da adoção de *marketing* de mercados-alvo.

Entre os 12 objetivos específicos estabelecidos pelo documento em discussão, para auxiliar a empresa no cumprimento de sua missão, estava o de "valorizar as ações de desenvolvimento de produtos e processos, de difusão de informação e de comercialização de tecnologias, serviços e produtos...".

Mais adiante, o referido documento estabelecia vários projetos gerenciais para auxiliar na consecução dos objetivos específicos já enunciados. Dentre os projetos, destacam-se três projetos da categoria Informação e Comunicação, os quais são: Comunicação na Embrapa (projeto 20), Sistema Embrapa de Informação (Projeto 2) e *Internet* (projeto 30). Embora seja possível estabelecer relações da *Web* com os três projetos, os dois últimos possuem ligações mais estreitas com a mesma. O projeto 2 - Sistema Embrapa de Informação - tinha como objetivo implementar um sistema informatizado e em rede que tornasse disponível todas as informações geradas pela empresa para a própria Embrapa e para o público em geral. O

projeto 30 - *Internet* - tinha como objetivo estabelecer uma política clara que regulasse a disponibilidade de informações institucionais e comerciais, padronizando formatos e estabelecendo mecanismos de proteção, em coordenação com o projeto Sistema Embrapa de Informação, com o projeto Comunicação na Embrapa e outros projetos ligados às relações comerciais e de cooperação da Empresa.

Já o documento Estratégia Gerencial da Embrapa, Macroprioridades/1997 (1997), estabelecia Inovar Métodos e Meios de transferência como prioridade institucional. Esta macroprioridade preconizava o seguinte:

> "... os avanços na área de telecomunicações e informática estão trazendo profundas modificações nos hábitos de informação do meio rural. As oportunidades de reciclagem dos produtores se ampliaram muito e se tornaram mais rápidas com a proliferação das antenas parabólicas e da comunicação via computador. Tais circunstâncias exigem que novos canais e processos de transferência de tecnologias sejam imaginados de sorte a melhorar a eficiência da Embrapa neste processo, aumentando a oferta de informações e o número de técnicos e produtores atendidos, a custos menores...".

Em 1998, após a etapa de redefinição de planos da Empresa, o documento Plano Diretor da Embrapa - Realinhamento Estratégico 1999-2003 (1998), reafirma todos os conceitos expostos anteriormente, apontando como fatores responsáveis por grandes transformações a globalização com abertura de mercado, a importância do meio ambiente, a reforma do Estado, a força do consumidor e a revolução tecnológica, caracterizando o agronegócio brasileiro do futuro como um setor competitivo, com qualidade e produtividade, tecnologicamente avançado, demandante de informação técnico-gerencial e promotor de emprego e renda.

A missão, agora redefinida, consiste em "...viabilizar soluções para o desenvolvimento sustentável[16] do agronegócio[17] brasileiro por meio de geração, adaptação e transferência de conhecimentos e tecnologias, em benefício da sociedade...". Importante caracterizar a visão do que seja cliente: "...a Embrapa considera como cliente todo o indivíduo, grupo ou entidade,

---

[16] Entende-se por desenvolvimento sustentável o arranjo político, sócio-econômico, cultural, ambiental e tecnológico que permite satisfazer as aspirações e necessidades das gerações atuais e futuras.

[17] O conceito de agronegócio engloba os fornecedores de bens e serviços à agricultura, os produtores agrícolas, os processadores, os transformadores e os distribuidores envolvidos na geração e no fluxo dos produtos agrícolas até o consumidor final. Participam também do agronegócio os agentes que coordenam o fluxo dos produtos, tais

pública ou privada, cujo sucesso em suas atividades dependa dos produtos e serviços, de natureza econômica ou social, oferecidos pela empresa e seus parceiros...".

Além disto, a Empresa coloca como objetivo que colaborará para a consecução de sua missão, o de viabilizar soluções tecnológicas para o desenvolvimento de um agronegócio competitivo em uma economia global, estabelecendo para isto que será necessário monitorar e analisar o comportamento dos mercados afins do agronegócio, identificando tendências e oportunidades.

Visando o fortalecimento de sua posição como instituição viabilizadora de soluções para o agronegócio e para consecução de políticas governamentais em bases sustentáveis e competitivas, o mesmo documento estabelece algumas diretrizes estratégicas para as atividades de pesquisa e desenvolvimento (P&D) e de transferência de tecnologia, entre outras. Reafirma uma política geral de administração embasada nos conceitos de *marketing* e qualidade total que assegure a) a disponibilidade de conhecimentos e tecnologias que tenham real interesse para a sociedade, b) a negociação para a distribuição de conhecimentos e das tecnologias gerados pela Empresa e por seus parceiros e c) a promoção dos conhecimentos e tecnologias entre os diversos públicos.

Especificamente para as atividades de pesquisa e desenvolvimento a Empresa pretende "...adotar mecanismos de levantamento e priorização de demandas e do grau de satisfação dos clientes com os conhecimentos e as tecnologias disponíveis...". Já a transferência de conhecimentos e tecnologias está apoiada nas políticas de negócios tecnológicos e de comunicação. Para isto, a Empresa deverá, entre outras ações, "...criar meios para a transferência de conhecimentos e tecnologias desenvolvidos ou mobilizados pela Embrapa para o maior número de clientes...". Além disso, a Embrapa "...implementará estratégias de comunicação empresarial que sejam orientadas à melhoria da interação interna e à potencialização do relacionamento da empresa com o ambiente externo, especialmente quanto a transferência de tecnologia...". Para atingir este objetivo, a Empresa deverá criar, manter e ampliar fluxos de comunicação de modo a estimular a interação entre a empresa e seus públicos interno e externo, além de inovar e modernizar os métodos e instrumentos de comunicação com os diversos públicos.

---

como o governo, os mercados, as entidades comerciais, financeiras e de serviços.

A ênfase de utilizar redes de computadores e redes de comunicação para transferir tecnologias fica mais clara quando, no mesmo documento, a empresa define projetos chamados estruturantes, em particular o projeto "Transferência de Tecnologia", cuja justificativa é a seguinte:

> "...os dias atuais exigem procedimentos mais ágeis de transferência tecnológica. Novos sistemas de comunicação e de informática permitem imprimir maior velocidade ao processo de transferência tecnológica. A empresa intensificará e institucionalizará novos mecanismos de transferência tecnológica, caracterizados pela rapidez com que a informação transita entre o pesquisador, as bases de dados e o usuário, pela rápida atualização das informações disponibilizadas, pelo estímulo à interação entre as equipes de pesquisadores, os agentes de assistência técnica e o usuário e pela facilidade de acesso às informações tecnológicas. Para tanto, as redes de computadores serão usadas para atendimento do público em geral e dos profissionais de assistência técnica que serão credenciados pela Embrapa...".

Todas estas políticas, diretrizes e objetivos apresentados possuem em comum o fato de atribuir grande importância à tecnologia da informação - especialmente os recursos de rede - como ferramenta indispensável ao processo de transferência de tecnologia, de soluções tecnológicas e da consolidação da imagem da Empresa junto aos seus públicos. Assim, são necessários estudos que identifiquem os mercados-alvo que possam alavancar o desenvolvimento tecnológico do setor rural a partir de uma maior - e mais direcionada - oferta de informações.

Além das intenções apresentadas, a empresa investiu pesadamente em recursos de informática na década de 90, implantando 37 redes locais, formadas com cerca de 5000 microcomputadores e mais de 200 workstations. Recentemente, concluiu projeto interligando suas 37 unidades descentralizadas em um sistema de comunicações via satélite, que possibilita o tráfego de som, dados e imagem (videoconferência).

A entrada da empresa na rede *Web* deu-se em 1995, quando foi implantada a estrutura de redes locais de computadores em todas as suas unidades e iniciou-se o processo de ligação destas redes locais na *Internet*. A interligação da totalidade da empresa foi um processo lento em razão das limitações impostas pela localização das unidades descentralizadas - geralmente

em áreas rurais. Apesar de ter adotado o protocolo padrão da *Internet* (TCP/IP), as soluções adotadas para elaboração da malha física da rede apresentam-se muito heterogêneas até a implantação e consolidação do *backbone* próprio, via satélite, em 1999.

No período de implantação e consolidação da rede, as unidades descentralizadas foram gradativa e isoladamente desenvolvendo suas páginas, valendo-se da maior ou menor capacidade técnica do pessoal da área de informática. Atualmente, todas as suas 37 unidades de pesquisa possuem páginas eletrônicas individualizadas, que são acessadas através de uma página principal, localizada na sede da empresa, em Brasília (http://www.embrapa.br), ou acessadas diretamente através de suas URLs individuais (*www.sigla-da-unidade.embrapa.br*). Por terem evoluído sob diferentes capacidades e visões, as páginas atualmente encontram-se muito diferenciadas nos aspectos de opções temáticas e de apresentação.

Não existe um padrão rigoroso entre as páginas das unidades, embora haja um documento oficial que estabelece cores, fontes, *links* padrões (conteúdo) e outros elementos a serem utilizados no seu processo de elaboração. Existem, também, diversas orientações quanto à responsabilidade pela elaboração e manutenção das páginas eletrônicas entre as unidades. Estima-se que a maioria das páginas foram criadas e estão sendo mantidas por técnicos da área de informática, com várias delas apresentando apenas informações institucionais

As páginas eletrônicas da unidade de pesquisa considerada começaram a ser disponibilizadas a partir do final de 1995 e atualmente publicam informações variadas sobre as atividades desenvolvidas, como projetos de P&D em andamento, equipe de pesquisadores, publicações *on-line*, serviços, produtos e etc. O *site* está acessível 24 horas/dia, sete dias/semana, sendo possível localizá-lo através dos mecanismos de busca mais conhecidos da *Web* como *altavista (*http://www.altavista.com*), yahoo (*http:www.yahoo.com*)* e *cade (*http://www.cade.com.br*)*. Consulta efetuada no *site* do Altavista no mês de julho de 2000, retornou um total de 55 páginas apontando para o *site* considerado.

3.3 O Estudo

O diagrama abaixo ilustra onde se situa a análise de *logs* no processo em investigação e quais seriam as suas aplicações práticas.

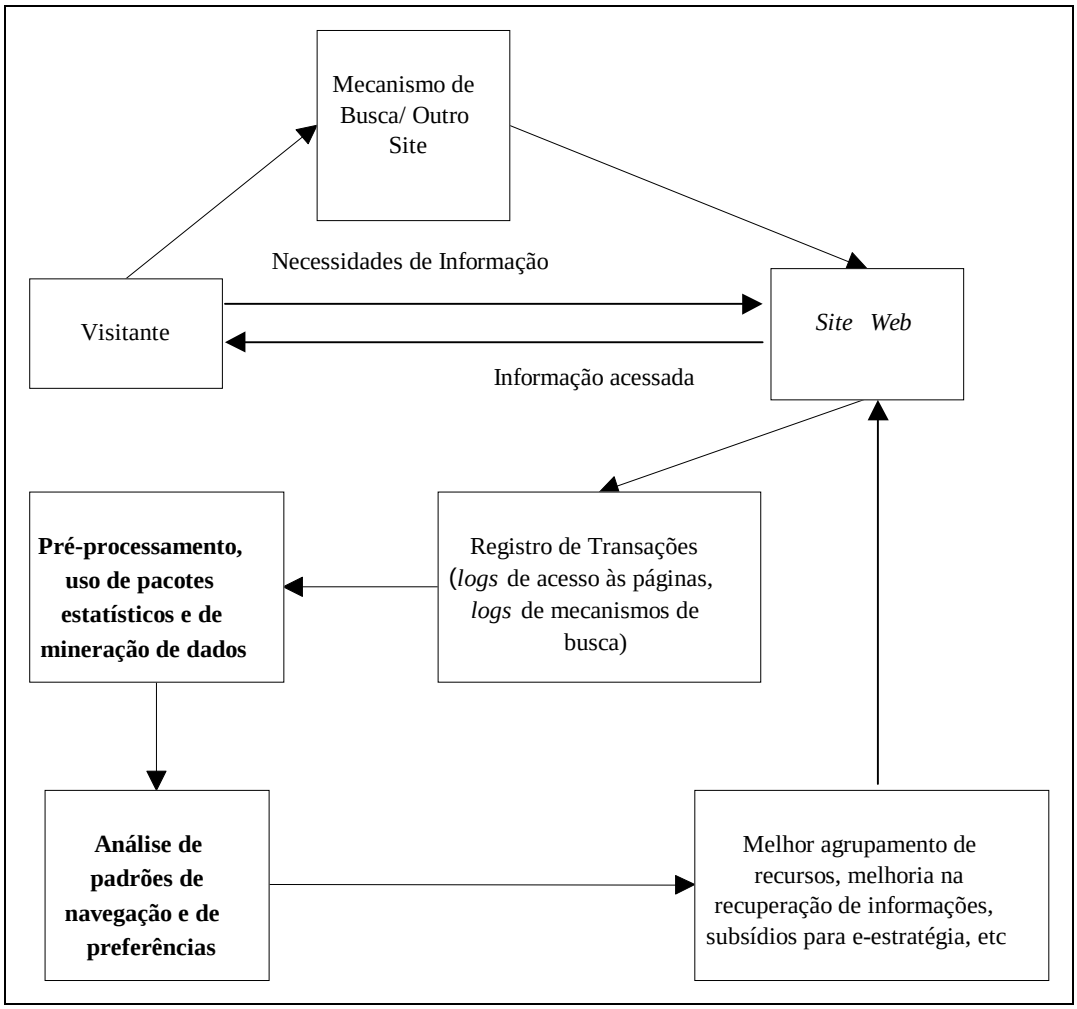

FIGURA 6 - Diagrama do escopo da pesquisa

Uma determinada necessidade por informação leva o visitante ao *site Web*. Uma vez no *site*, o visitante segue trilhas de navegação guiadas pelas suas preferências e necessidades por informação. As trilhas percorridas através dos conteúdos/recursos do *site* são automaticamente registradas nos arquivos de *logs*. As "pegadas" deixadas são preparadas (pré-processadas) e, após agrupadas, são submetidas a procedimentos estatísticos tradicionais e de mineração de dados. Os resultados destas análises indicarão padrões e tendências que poderão ser considerados na definição, na estruturação e no agrupamento dos conteúdos e

recursos disponibilizados pelo *site Web*. O resultado esperado será o de proporcionar maior "fluxo" ao visitante que acessa o *site*. Por "fluxo" entenda-se o processo de experiência ótima, proporcionado pelo balanceamento entre desafios e habilidades requeridas aos humanos na interação com um sistema de *hypermidia* (CSIKSZENTMIHALYI, 1977; CSIKSZENTMIHALYI e LEFEBRE, 1989; CSIKSZENTMIHALYI, 1990, citados por HOFFMAN e NOVAK, 1995).

3.3.1 Descrição dos dados

a) conjunto de dados referente às sessões realizadas no *site*:

Foram coletados registros de acessos ao *site* no período de 02/02/1999 a 30/04/2000 (453 dias). O registro é realizado de forma automática pelo servidor HTTP no arquivo padrão *access_log*, formato ASCII (txt) , sendo composto dos seguintes campos (W3C, 1999):

computador-remoto - usuário - data:hora "método/requisição" status bytes-transferidos

Exemplo de linhas do *log* do servidor httpd, no formato "common":

```
...
scm619.ufrgs.br - - [30/Jun/1999:20:40:01 -0300] "GET /index.html HTTP/1.0" 200 15300
scm619.ufrgs.br - - [30/Jun/1999:20:42:09 -0300] "GET /images/logotipo.gif HTTP/1.0" 200 10234
scm619.ufrgs.br - - [30/Jun/1999:20:44:09 -0300] "GET /unidade/equipe.html HTTP/1.0" 200 37285
...
```

- computador-remoto - nome do computador remoto que está requisitando o recurso (ou número IP, se o nome não estiver disponível)
- data:hora - data e hora da requisição;
- método e requisição - método da requisição e recurso (arquivo html, doc, pdf, cgi, gif, jpg, etc.) requisitado pelo computador remoto, na exata forma em que foi digitada pelo requisitante;
- status - código de status do HTTP referente à transação. Os status mais comuns são:

  200 - recurso foi transferido com sucesso
  206 - transferência foi interrompida
  304 - recurso requisitado foi servido pela *cache* do requisitante
  404 - recurso não foi encontrado
- bytes-transferidos - número de bytes transferidos

Um resumo das transações registradas no arquivo original, durante o período amostrado, pode ser visualizado na tabela abaixo (Tabela 1):

TABELA 1 - Resumo das transações registradas no arquivo original

Item do *log*	Total de linhas	(%)
Requisições para páginas HTML (*pageviews*)	52.717	12,8
Uso do mecanismo de busca do *site* (cgi)	5.234	1,3
Uso do mecanismo de busca do informativo externo (cgi)	1.783	0,4
Cadastro de leitor do informativo externo (cgi)	161	0,04
Transferência do Livro de Receitas de Carne (zip)	325	0,08
Requisições para arquivos gráficos (.jpg, .gif)	325.554	78,8
Outras requisições* (status diferente de 200 ou 304, falhas no sistema de logs, outras páginas, cgi´s ou scripts de pouco interesse)	27.608	6,7
Total de registros no Log (*hits*)	413.382	100

b) Conjunto de dados referente às palavras-chave inseridas no mecanismo de busca do *site*:

Foram coletadas, analisadas e classificadas palavras-chave inseridas no mecanismo de busca do *site* entre às 18h25 de 08/11/1999 e 11h09 de 15/06/2000. A amostra totalizou 2.905 termos para busca coletados em 1.473 sessões. Neste conjunto de dados não foi aplicado nenhum critério de seleção ou amostragem, considerando-se todas as consultas realizadas no *site* no período mencionado.

O registro é realizado de forma automática por um programa no servidor HTTP em arquivo formato ASCII (txt) , sendo composto pelos seguintes campos:

computador-remoto - data:hora - “palavra(s)-chave“ - bytes-transferidos - num. doc. recuperados

Exemplos de linhas do *log* de consultas:

```
...
attila.urcamp.tche.br [03/Jan/2000:15:30:13] "pastagem" 80003 45
attila.urcamp.tche.br [03/Jan/2000:15:42:09] "pastagem cultivada" 42302 20
proxie.unesp.br [03/Jan/2000:15:43:44] "a historia da pecuaria" 154020 189
...
```

- computador-remoto - nome do computador remoto em que está sendo feita a consulta (ou número IP, se o nome não estiver disponível)
- data:hora - data e hora da consulta;
- bytes-transferidos - número de bytes transferidos
- número de documentos recuperados

As consultas foram importadas para um sistema de banco de dados, consistidas e após classificadas conforme descrito no Capítulo 4.

3.3.2 Limites e pressupostos

a) Quanto à origem e identificação dos acessos

É importante fazer uma observação sobre o dado *computador-remoto*. Como a maior parte dos conteúdos são disponibilizados na *Web* sem a necessidade de *login/password* para acesso, identificar um visitante a partir da origem (computador-remoto) de uma requisição, procurando dotá-la de características individuais, constitui-se em uma tarefa de resultados probabilísticos. Várias são as causas que podem esconder ou mascarar o visitante que está atrás de uma sessão, tais como:

a) o equipamento requisitante pode utilizar um *proxie* ou um provedor para acessar a rede e, neste caso, é o nome deste último que aparece nos *logs*;
b) o sistema de DNS (Domain Name Service), que é responsável pela resolução de nomes de computadores a partir dos seus endereços IP, pode falhar; neste caso não será possível saber domínio da origem;

c) várias pessoas podem utilizar um mesmo computador ao longo do tempo;
d) várias pessoas podem utilizar um mesmo endereço ao mesmo tempo;
e) as requisições podem estar sendo feitas por um *robot* ou *spider* e, neste caso, as sessões retratam uma lógica diferente da lógica humana;

Uma das soluções encontradas para o rastreamento de usuários individuais são os *cookies*. Entretanto, este sistema também não se apresenta confiável. Além de possuírem validade determinada (tempo de expiração), nem sempre os usuários - em razão de privacidade - permitem a sua utilização.

Desta forma, a tarefa de tentar identificar visitantes individuais tem sido apresentada na literatura como um conjunto de heurísticas que, pela utilização de uma combinação de métodos, apenas diminuem o erro no processo. No caso desta investigação, onde o interesse recaiu sobre a origem da visita e não sobre o visitante em sí, apenas o item b, acima, foi contornado, através de pós-processamento dos *logs* pelo programa *logresolve*, disponível para o servidor HTTP Apache[18].

b) quanto à estruturação das sessões e visitantes

Não existe muito consenso na literatura a respeito da diferença entre os termos *sessão* e *visita*. NOVAK e HOFFMAN (1996) discriminam estes termos considerando *sessão* como uma série de *visitas* consecutivas feitas por um usuário em uma série de *sites Web*, em um intervalo de tempo contínuo. Talvez esta abordagem seja mais adequada para pesquisadores que investiguem o comportamento de usuários em um ambiente controlado como, por exemplo, aplicações que registrem as ações de usuários através de programas executados nos clientes (*client-side*) e não nos servidores (*server-side*). Já grande parte da literatura não faz distinção entre os termos *sessão* e *visita*, havendo mesmo a predominância do primeiro. Para o propósito deste trabalho os dois termos são equivalentes.

Existem basicamente dois critérios para agrupar as entradas do arquivo de *log* em sessões: uma nova sessão começa quando: 1) a duração da sessão excede um determinado limite de

[18] www.apache.org, consulta em 04/2000.

tempo (COOLEY *et al.*, 1997) , ou 2) o tempo decorrido entre dois acessos consecutivos excede um determinado limite de tempo (COOLEY *et al.*, 1997; CHEN *et al.*, 1996). Neste trabalho optou-se por utilizar a última abordagem, estabelecendo o limite de tempo máximo entre dois acessos consecutivos em 30 minutos. Após este tempo a sessão foi desligada e o próximo acesso foi considerado uma nova sessão.

Abaixo, encontra-se a representação do algoritmo utilizado para a identificação das sessões, o qual gera um número serial único para cada uma das sessões:

```
ordene arquivo por origem da requisição, por data e por hora do  acesso
leia registro
enquanto nao é fim do arquivo
  origem_memória = origem_disco
  enquanto nao é fim do arquivo  e   (origem_memória = origem_disco)
     número_da_sessão = número_da_sessão + 1
     hora_do_último_acesso = hora_do_acesso
     enquanto nao é fim do arquivo  e (origem_memória = origem_disco ) e (hora_do_acesso - hora_do_último_acesso ) < 30 min.
        grave número_da_sessao
        hora_do_último_acesso = hora_do_acesso
        leia registro
     }
   }
}
fim
```

Por *visitante* será entendido o indivíduo que realiza uma sessão no *site*. No caso deste estudo, o visitante não será identificado (NOVAK e HOFFMAN, 1996). Será, também, utilizado o termo *pageview* para designar o acesso a um documento HTML do *site*.

Assim, neste estudo será usado o termo ***sessão*** para designar um conjunto de ***pageviews*** requisitadas por um **visitante não identificado,** cuja seqüência só é quebrada se o tempo entre uma *pageview* e outra quaisquer ultrapassar 30 minutos.

c) quanto ao tempo de exposição de páginas e do *site*

Já o tempo de exposição de cada página foi calculado utilizando-se a fórmula (ZAÏANE, 1998):

tempo_de_exposição_da_página = hora_do_acesso - hora_do_acesso_posterior

Neste modelo não é possível obter o tempo de exposição de sessões de um único acesso, bem como do último acesso de cada sessão. Logo, o total de segundos da sessão - ou o tempo de exposição do *site* - será a soma de segundos das ***n-1*** páginas acessadas na sessão, sendo ***n*** maior do que 1.

d) quanto à ação de *robots* e *spiders*

Após o primeiro pré-processamento o arquivo foi submetido a procedimentos de estatísticas descritivas. A análise dos resultados revelou alguns comportamentos de navegação que distorciam medidas, tais como, o número de acessos em uma única sessão e o número de sessões de um mesmo visitante ao longo do tempo.

Uma verificação mais detalhada mostrou que aqueles acessos eram oriundos de dispositivos automáticos de indexação (*robots/spiders*) dos principais diretórios de busca da *Web*. ZAÏANE (1998) sugere que estes acessos sejam mantidos apenas quando o objetivo é o estudo do comportamento destes agentes no *site*.

Foram identificados 6.200 (10,3%) acessos destes dispositivos. Pelo fato de serem agentes automatos, cujo comportamento de navegação está determinado por algoritmos que visam diferentes objetivos e seguem lógicas igualmente distintas e fracamente dirigidas pela lógica humana, estes acessos poderiam comprometer algumas conclusões do estudo. Entretanto, em razão do formato do *log* ser o COMMON e não o EXTENDED, não foi possível identificar os acessos de robots via reconhecimento dos agentes. Esta identificação ocorreu através de exame visual e do reconhecimento dos nomes de domínios. Logo, no conjunto de acessos cujo domínio não foi resolvido poderia, ainda, conter outros acessos daqueles agentes.

d) Outros aspectos

Do total de 18.047 sessões, aquelas com somente um acesso (9.154) representaram, também, um fator de distorção das medidas e revelaram ser de pouca contribuição para o estudo do comportamento de navegação do visitante durante uma sessão. Da mesma forma as sessões que consistiam de uma mesma página requisitada *n* vezes (239).

Além disso, muito embora os acessos de fora do país representassem 7.274 requisições, mais da metade (52%) era oriunda de *robots*. Dos restantes, 32% eram sessões de apenas um acesso. Aliado ao fato de que o *site* não possui opções para idiomas, estes dados colocaram em dúvida a necessidade de manter aqueles acessos no conjunto de dados final.

3.3.3 Limpeza dos dados e definição da amostra

Do conjunto de *pageviews* inicialmente considerado, foram selecionados os acessos aos arquivos de conteúdo (HTML), os acessos aos dois mecanismos de busca, o cadastro de leitor e a transferência do livro de receitas de carne cujo status de transferência fosse 200 (página transferida com sucesso) ou 304 (página requisitada mas servida pela cache do *browser* requisitante), o que totalizou um conjunto inicial de 60.220 observações (14,5%). Utilizando-se de softwares específicos, os acessos foram então agrupados por sessão.

Pelas razões já descritas no item 3.3.2 e, uma vez que este estudo discutiria também o comportamento de navegação do usuário no *site*, optou-se por diminuir as distorções citadas retirando do arquivo os acessos de *robots*/*spiders*, de endereços com domínios não resolvidos, de endereços de fora do país e de sessões com somente um acesso.

Após estas análises iniciais foi possível definir melhor a amostra do estudo como:

> **registro de páginas transferidas com sucesso para visitantes não identificados, oriundos do domínio '.br' , que acessaram mais de uma página - diferentes - durante a visita, no período de 2 de fevereiro de 1999 a 30 de abril de 2000.**

Na página seguinte é apresentada uma tabela comparativa entre o arquivo original e a amostra selecionada (Tabela 2):

TABELA 2 - Comparação entre o arquivo original e a amostra selecionada

	Arquivo Original	**Amostra**	**(%)**
Total de requisições consideradas	60.220	26.961	44,8
Requisições para páginas HTML (*pageviews*)	52.717	22.641	43,0
Uso do mecanismo de busca do *site* (cgi)	5.234	3.055	58,4
Uso do mecanismo de busca do informativo externo (cgi)	1.783	959	53,8
Cadastro de leitor do informativo externo (cgi)	161	91	56,5
Transferência do Livro de Receitas de Carne (zip)	325	215	66,2
Acessos de fora do País	7.274	0	0,0
Acessos com origem desconhecida (endereço IP não resolvido)	19.778	0	0,0
Acessos de Robots/Spiders conhecidos	6.200	0	0,0
Sessões com somente um *pageview*	9.154	0	0,0
Sessões cuja seqüência de requisições se resumia a somente uma página do *site*	239	0	0,0
Megabytes transferidos	690	318	46,1
Número de sessões	18.047	4.729	26,2
Computadores diferentes	9.445	3.488	36,9
Domínios de terceiro nível diferentes	1.096	684	62,4

O percentual de sessões incluídas na amostra em relação ao arquivo original ficou em 26,2%. Entretanto, estas sessões foram responsáveis por 44,8% das requisições previamente consideradas. Da mesma forma, a amostra reteve 62,4% dos domínios de terceiro nível[19] e 36,9% dos computadores diferentes, tendo estes sido responsáveis por 46,1% do volume transferido. Com base nestes dados, considerou-se a amostra significativa para os objetivos do estudo.

[19]Diz-se “domínio de terceiro nível” a parte do endereço em que, geralmente, aparece o nome da organização. No caso de instituições de ensino e pesquisa, o nome geralmente está no segundo nível. Estes foram tratados, entretanto, como de terceiro nível, de forma a padronizar a análise.

CAPÍTULO 4: Resultados e Discussão

Após a limpeza dos dados, agrupamento em sessões e definição da amostra, as sessões restantes foram submetidas a programas para extração de estatísticas descritivas e para mineração de dados. A fim de se obter maior flexibilidade para análise dos caminhos dos visitantes, os dados das sessões foram, também, importados para um software dotado de uma linguagem *SQL-like*, o que permitiu a elaboração de consultas de forma *ad-hoc*.

Os resultados serão apresentados e discutidos, considerando:

a) as estatísticas gerais de acesso ao *site;*
b) as preferências e padrões primários de navegação dos visitantes no *site;*
c) as preferências explícitas dos visitantes, simbolizadas pelos termos inseridos no mecanismo de busca do *site.*

Ao final, procurar-se-á elaborar um perfil do uso do *site* e de suas particularidades, considerando as análises e discussões sobre os três aspectos enfocados.

4.1 Análise e discussão das estatísticas gerais de acesso ao *site*

Após definida a amostra foram realizadas algumas análises das sessões visando apresentar um resumo descritivo dos dados em seus aspectos mais elementares. Os resultados destas análises estão apresentados abaixo.

Dos 453 dias de acompanhamento o *site* registrou acessos externos em 416 (91,6%). A média do número de sessões por dia foi de 11,3. Ressalte-se que a amostra selecionada corresponde a apenas 26,2% das sessões, conforme já visto na Tabela 2.

A Tabela 3 mostra que 37,1% das sessões não iniciaram pela página principal. Vários fatores podem levar a isto, dentre os quais estão os acessos realizados a partir de mecanismos de buscas, acessos à páginas *linkadas* por outro *site* e a entrada direta do nome do documento que pode ser feita pela digitação da URL inteira ou pelo uso de *bookmarks* pelo visitante. Os dados mostram que as sessões iniciadas pela página principal (*home page*) requisitaram 1,3 páginas a mais em relação as sessões iniciadas em outras páginas do *site*, muito embora estas últimas possuissem um tempo médio por *pageview* maior. Em média, o visitante requisitou 5,7 páginas, tendo ficado conectado aproximadamente 8:36 minutos no *site*.

TABELA 3 - Características das sessões considerando sua página inicial

Sessão	**Número de sessões**	**%**	**Média de Pageviews por Sessão**	**Tempo Médio por pageview em Segundos**	**Tempo Médio por sessão em segundos**
Iniciada na Home-Page	2.973	62,9	6,2	86,5	534,1
Iniciada em outras páginas	1.756	37,1	4,9	99.0	484,5
Totais	4.729	100,0	5,7	90,5	515,7

A distribuição das sessões por sua origem, considerando os diferentes domínios de segundo nível pode ser visualizada na Tabela 4. Observa-se que 71,9% dos acessos foram provenientes do domínio ".com.br", e 22,3% originados em instituições de ensino e pesquisa (predominantemente universidades). Procurou-se dividir esta última categoria considerando a sua localização, classificando-se separadamente as universidades particulares e instituições públicas do RS, a fim de ter uma idéia mais aproximada do uso do *site* por instituições mais aderentes geograficamente. Os acessos de provedores de rede (.net), órgãos do governo (.gov) e outros representaram apenas 5,8%.

A média de *pageviews* por sessão mostrou pequenas diferenças, tendo as instituições de ensino e pesquisa do RS apresentado um número ligeiramente mais elevado (6,4). Estas, juntamente com as instituições de ensino e pesquisa federais e de outros Estados, apresentaram também um maior tempo de conexão ao *site* (11 min.) Uma análise mais detalhada, considerando o domínio de terceiro nível, encontra-se nas Tabelas 5, 6 e 7.

TABELA 4 - Distribuição do número de sessões por origem

Domínio	**Número de Sessões**	**%**	**Número *pageviews***	**Média de pageviews p/sessão**	**Tempo Médio de exposição do *site***
Organizações Comerciais/ISPs	3.399	71,9	19.215	5,7	466,8
Instit. de Ensino e Pesquisa	937	19,8	5.514	5,9	671,8
Provedores de Rede	149	3,2	807	5,4	590,6
Instit. de Ensino e Pesquisa do RS	113	2,4	718	6,4	673,3
Órgãos do Governo	98	2.1	583	6,0	457,0
Outros	33	0,6	124	3,8	415,1
Totais	4.729	100,0	26.961	5,7	515,7

A Tabela 5 mostra com mais detalhes os acessos realizados por organizações que fazem parte do domínio ".com.br", as quais realizaram mais de 25 sessões no *site*. As 19 organizações listadas abaixo, que representam apenas 2,8% do total de domínios de terceiro nível registrados, foram responsáveis por 38,9% das sessões realizadas no *site*. Nota-se que existe uma predominância de provedores que servem a Região Sul e Sudeste do Brasil, chamando a atenção o número de sessões com origem na Acessonet. Uma análise mais detalhada mostrou que o interesse dos usuários daquele provedor era o *link* de receitas de carne do *site*. Das 572 sessões daquela origem, 512 (89,5%) acessaram o *link*. A acessonet provê acesso para o UOL, estando seus usuários concentrados predominantemente em São Paulo.

O número médio de *pageviews* requisitado por cada um dos domínios de terceiro nível varia consideravelmente, sendo um dos destaques a Procergs (Companhia de Processamento de Dados do Estado do Rio Grande do Sul).

TABELA 5 - Distribuição da origem dos acessos por domínio de terceiro nível, considerando apenas organizações com 25 sessões ou mais, do domínio comercial ".com.br"

Domínio de terceiro nível	**Número de Sessões**	**Média de pageviews p/sessão**
acessonet	572	4,3
zaz	366	4,9
alternet	147	4,4
procergs	130	6,8
nutecnet	115	4,7
matrix	74	5,9
conex	67	5,8
uol	46	3,7
onda	41	6,3
estaminas	31	3,5
sercontel	30	7,7
cultura	29	3,3
sti	29	3,0
homeshopping	28	5,2
svn	27	3,6
horizontes	27	3,3
viavale	26	8,5
ez-poa	26	5,2
mandic	26	3,7

Total (19)	1.837	4,9

A Tabela 6 apresenta uma listagem das sessões realizadas no *site* nas quais a origem foram instituições de ensino e pesquisa federais ou localizadas fora do Estado do RS, que realizaram mais de 10 sessões no *site*. As instituições listadas representam 2,6% do total de domínios de terceiro nível diferentes registrados, sendo responsáveis por 18,6% do total de sessões realizadas.

Nota-se, entretanto, que 48% dos acessos foram originados por outras unidades da Embrapa. No restante dos acessos novamente aparece a predominância de instituições localizadas na Região Sul e Sudeste, com algumas poucas sessões, também, do Nordeste (UFPE, UFBA). Supõe-se que a afinidade com o tema do *site* seja por possuírem cursos fortes na área de ciências agrárias e/ou, também, por pertencerem a regiões onde a pecuária ocupa lugar de destaque na economia.

O número médio de *pageviews* por sessão também varia bastante entre as instituições listadas, não sendo surpresa que a UFSM (Universidade Federal de Santa Maria) e a UFRGS (Universidade Federal do Rio Grande do Sul) apresentaram uma média elevada de páginas transferidas por sessão, dada a grande tradição de ensino e pesquisa na área de ciências agrárias e também por possuírem estreitas ligações com a empresa que mantém o *site*, seja por atividades de pesquisa conjunta, ou pelo fato da formação dos pesquisadores em sua maioria ter se dado naquelas instituições.

TABELA 6 - Distribuição da origem dos acessos por domínio de terceiro nível, considerando apenas instituições de ensino e pesquisa federais e/ou de fora do RS, com 10 sessões ou mais

Domínio de terceiro nível	UF	Número de Sessões	Média de pageviews p/sessão
embrapa	(*)	422	3,9
ufsm	RS	65	6,6
ufrgs	RS	63	6,6
pop-mg	MG	44	3,6
rct-sc	SC	29	5,9
usp	SP	26	4,3
ufpe	PE	25	3,6
unesp	SP	25	4,2
ufsc	SC	24	7,3
pop-rs	RS	24	7,3
intranetparana	PR	22	5,2
ufmg	MG	21	2,8
ufba	BA	18	8,7
ufu	MG	18	5,1
unicamp	SP	16	2,9
redeminas	MG	16	3,0
ufla	MG	11	4,6
ufrrj	RJ	10	3,8
Total (18)		879	5,0

(*) A Embrapa possui unidades em grande parte dos Estados do País

Já na Tabela 7 estão listadas as instituições de ensino e pesquisa do RS, as quais realizaram mais de 10 sessões no *site*. Estas representaram apenas 0,6% do total de domínios de terceiro níveis diferentes, mas que realizaram 2% das sessões totais no *site*.

A URCAMP (Universidade da Região da Campanha), cuja região de influência está inserida na região de abrangência da unidade de pesquisa que mantém o *site*, realizou 45,8% das sessões. Entretanto, a UNISC (Universidade de Santa Cruz), juntamente com a UCPEL (Universidade Católica de Pelotas) foram as que mais requisitaram páginas por sessão (7,0 e 6,4 respectivamente.).

TABELA 7 - Distribuição da origem dos acessos por domínio de terceiro nível, considerando apenas instituições de ensino e pesquisa do RS, com 10 sessões ou mais.

Domínio de terceiro nível	Número de Sessões	Média de pageviews p/sessão
Urcamp	44	4,2
Upf	21	4,8
Ufpel	20	6,4
Unisc	11	7,0
Total (4)	96	5,6

Na Tabela 8 pode-se visualizar a evolução dos acessos por mês ao longo do período de coleta dos dados. Nota-se um forte incremento no número de sessões realizadas a partir do mês de maio de 1999. Desconfia-se que um maior número de acessos em determinado mês pode estar associado ao calendário das universidades que mantém cursos na área de ciências agrárias.

Faz-se necessário registrar que a queda no número de sessões nos meses de julho e agosto de 1999 podem estar associadas à troca de *backbone* do *site*. Em meados de julho o *link* da Rede Tche, que ligava o *site* à *Internet*, foi substituído pela ligação da rede privativa de satélite da Empresa. Esta troca provocou alguns problemas de roteamento, colocando o *site* temporariamente "invisível" para algumas partes da *Internet*. Isto também pode explicar o repentino aumento no número médio de *pageviews* por sessão, bem como uma diminuição no tempo médio de exposição por página e do *site*. Ou seja, com a conexão do *site* na *Internet* seis vezes mais rápida e com menos sessões, as páginas foram transferidas mais rapidamente e, provavelmente, motivaram os visitantes a transferir mais páginas.

De fato, a Tabela 8 mostra que a partir de setembro de 1999 os tempos de exposição e o número de *pageviews* por sessão se estabilizaram visivelmente.

TABELA 8 - Distribuição do número de sessões realizadas por mês no s*ite*

Mês/ Ano	**Número de sessões**	**Média de Pageviews p/sessão**	**Tempo Médio de exposição/ página**	**Tempo Médio de exposição do *site***
Fev/1999	63	5,7	135,6	773,0
Mar	170	5,9	114,7	677,0
Abr	188	6,4	145,4	930,8
Maio	304	5,7	104,9	597,9
Jun	372	5,7	109,7	625,3
Jul	245	7,0	80,8	565,3
Ago	280	7,4	79,0	584,6
Set	391	5,5	80,9	444,7
Out	416	5,3	86,5	458,7
Nov	476	4,5	84,9	382,0
Dez	265	5,3	80,8	428,5
Jan/2000	286	5,8	86,5	501,5
Fev	383	5,8	81,0	469,9
Mar	383	5,9	80,1	472,9
Abr	507	5,3	84,9	450,1

Já a Tabela 9 apresenta a distribuição das sessões considerando os dias da semana. Nota-se uma queda de cerca de 40% em média no número de sessões realizadas nos finais-de-semana, o que já era esperado. O tempo de exposição também diminui sensivelmente, possivelmente em razão de menor atividade na rede, possibilitando a transferência das páginas mais rapidamente. O número de *pageviews*, no entanto, não apresenta grande diferença em relação ao resto da semana. Como era de se esperar, 95,2% das sessões realizadas nos finais-de-semana foram originadas no domínio ".com.br".

TABELA 9 - Distribuição do número de sessões por dia da semana no *site*

Dia da semana	Número de sessões	%	Média de pageviews por sessão	Tempo médio por pageview em segundos	Tempo médio por sessão em segundos
Dom	485	10,3	6,0	73,3	442,1
Seg	792	16,8	5,9	88,5	523,1
Ter	746	15,8	5,8	96,5	560,6
Qua	773	16,4	5,6	91,8	513,3
Qui	791	16,7	5,3	102,4	546,8
Sex	677	14,3	5,8	91,2	535,7
Sab	465	9,8	5,6	80,3	450,9

A distribuição do número de sessões ao longo das horas do dia é apresentada na Tabela 10. Ela mostra que o pico ocorre geralmente das 15 horas às 15h59 da tarde, mas também permanece alto das 14 horas às 16h59. Pela manhã o número de sessões já aumenta a partir das 9 horas. Já a média de *pageviews* por sessão permanece sem grandes alterações ao longo do dia. Os tempos de exposição, todavia, variaram consideravelmente, não ficando claro o motivo.

TABELA 10 - Distribuição das sessões por hora do dia no *site*

Hora	Número de sessões	Média de Pageviews por sessão	Tempo Médio por pageview em segundos	Tempo Médio por sessão em segundos
0	149	5,7	92,7	532,6
1	85	5,5	86,2	473,9
2	36	5,2	75,2	388,5
3	17	6,1	102,8	628,6
4	16	5,1	111,0	568,8
5	14	6,3	54,9	371,6
6	12	6,7	106,1	707,3
7	62	6,1	78,6	495,1
8	142	6,3	95,8	606,9
9	263	5,5	92,4	505,7
10	303	5,6	95,3	542,8
11	297	5,2	88,4	457,3
12	262	6,1	91,3	558,2
13	280	5,9	94,4	552,1
14	304	5,8	103,7	610,3
15	364	6,0	109,0	654,3
16	327	5,2	84,9	445,3
17	271	5,6	99,6	561,1
18	239	5,5	86,6	476,6
19	270	5,5	77,8	429,2
20	268	5,7	87,2	504,1
21	296	5,9	82,0	487,6
22	271	6,0	76,4	464,7
23	181	5,5	74,1	407,7

Não apresentando-se o tempo médio de duração das sessões como uma distribuição normal, possuindo alto desvio padrão e grande assimetria, as sessões foram agrupadas em intervalos, considerando as classes mostradas na Tabela 11. Esta distribuição em classes mostra que apenas 26,9% das sessões tiveram duração superior a 10 minutos.

TABELA 11 - Distribuição das sessões no *site* segundo categorias de duração total

Intervalo em segundos	Número de sessões	%	% acumulado
0 - 120	1350	28,5	28,6
121 - 240	882	18,7	47,2
241 - 360	561	11,9	59,1
361 - 480	373	7,9	67,0
481 - 600	294	6,2	73,1
601 - 720	226	4,8	77,9
721 - 840	172	3,7	81,6
841 - 960	145	3,1	84,7
>960	726	15,4	100,0

Da mesma forma, a distribuição do número de sessões em categorias de números de *pagreviews* por sessão revelou-se muito assimétrica, como mostra a Tabela 12. Nota-se que a grande maioria das sessões (73,0%) era constituída por 1 a 6 páginas. Um número expressivo de sessões, entretanto, transferiu mais de 15 páginas (5%).

TABELA 12 - Distribuição das sessões segundo categorias de número de *pageviews* por sessão

Números de páginas	Número de sessões	%	% acumulado
1 - 3	2066	43,7	43,7
4 - 6	1387	29,3	73,0
7 - 9	589	12,5	85,5
10 - 12	282	6,0	91,5
13 - 15	165	3,5	95,0
>15	240	5,0	100,0

4.2 Análise e dicussão dos caminhos dos usuários ao atravessar o *site*

Utilizando-se as 2.973 sessões que iniciaram a página principal, procurou-se saber quais eram as opções do visitante ao entrar no *site*. As Figuras 7 a 12 são representações dos caminhos tomados pelos visitantes ao escolher os principais *links* da página.

Os diagramas são apresentados em duas versões: uma considerando a primeira escolha do usuário e outra considerando as mesmas escolhas na sessão como um todo, desconsiderando a ordem em que foram feitas. A notação **A-B*** significa que o visitante iniciou sua sessão na página **A**, passou diretamente ("**-**") para a página **B**, continuando sua sessão através do *site*, ou não ("*****"). Já á notação **A*B*** significa que o visitante iniciou sua sessão na página **A**, acessou - ou não - outras páginas ("*****") antes de chegar a página "**B**", continuando sua sessão através do *site*, ou não ("*****").

A Figura 7 mostra a primeira opção dos usuários ao entrar na página principal do *site*. Nota-se que mais de 45% do primeiro *click* recai sobre os *links* "Índice de Atividades de Pesquisa" (17%), "Publicações" (14,7%) e "Serviços" (13,7%), revelando certa objetividade dos visitantes em saber o que a instituição está fazendo e o que ela tem para oferecer.

A análise do *link* sobre "Tecnologias, Serviços e Produtos", em particular, pode revelar a necessidade de ações que visem a melhoria do atendimento e/ou a maior disponibilização de informações ao público.

Por outro lado, o uso do "mecanismo de busca" do *site* também é representativo (11,5%), ou seja, praticamente 1 em cada 10 sessões inicia naquele *link*. Uma análise mais detalhada mostrou que 71,1% dos termos utilizados para busca estavam relacionados com o conteúdo oferecido pelo *site*, sendo 64,2% sobre assuntos diretamente ligados as atividades de P&D da unidade investigada, 4,4% sobre questões administrativas e 2,5% sobre receitas de carne bovina e ovina. Ressalta-se que a análise dos termos utilizados nas consultas foi efetuada utilizando-se critérios de amostragem diferentes dos utilizados para a amostragem das sessões. Para maiores informações sobre este tópico, consultar a sessão 4.3.

O acesso ao *link* "Novidades" também não deixa de ser representativo, revelando que um

determinado percentual dos visitantes (8,8%) tem um comportamento prospectivo e, provavelmente, esperava encontrar conteúdo de promoção atualizado no *site*. O restante das escolhas consideradas possuem conteúdo mais estático e de natureza informativa sobre o funcionamento da empresa.

Visando encontrar regras de preferência considerando o domínio de segundo nível (comercial, governo, instituições de ensino, etc) os dados da Figura 7, em particular, foram testados utilizando-se um software de classificação e geração de regras (QUINLAN, 1993). As regras geradas pelo software estão apresentadas na Tabela 13.

TABELA 13 - Regras produzidas considerando o conteúdo preferido na primeira escolha e o domínio de segundo nível da origem das sessões

Conteúdo preferido na primeira escolha	**Número de casos**	**Erros**	**(%) Erro**	**(%) Confiança**	**Classe**
Pesquisa	506	132	26,1	72,4	Comercial
Serviços	407	90	22,1	76,3	Comercial
Mecanismo de busca	343	72	21,0	77,3	Comercial
Novidades	261	63	24,1	73,8	Comercial
Informações sobre a Unidade	199	45	22,6	75,0	Comercial
Informações sobre a Equipe Técnica	181	77	42,5	54,6	Instituição de Ensino e Pesquisa

A Tabela 13 diz que havia 72,4% de confiança de que uma sessão, a qual a primeira escolha foi o conteúdo ‘Pesquisa”, era da classe “comercial”. É notória a predominância desta classe nas regras geradas. A preferência pelo conteúdo “Equipe Técnica” na primeira escolha, entretanto, foi atribuída à classe “Instituição de Ensino e Pesquisa”. Uma análise mais detalhada mostrou que a maior parte destes acessos (66 de 181 casos) era originária de outras unidades da própria empresa que mantém o *site*, fato que certamente influenciou na sua elaboração. Salienta-se que as regras geradas não foram testadas contra um conjunto de dados para teste (*test set*).

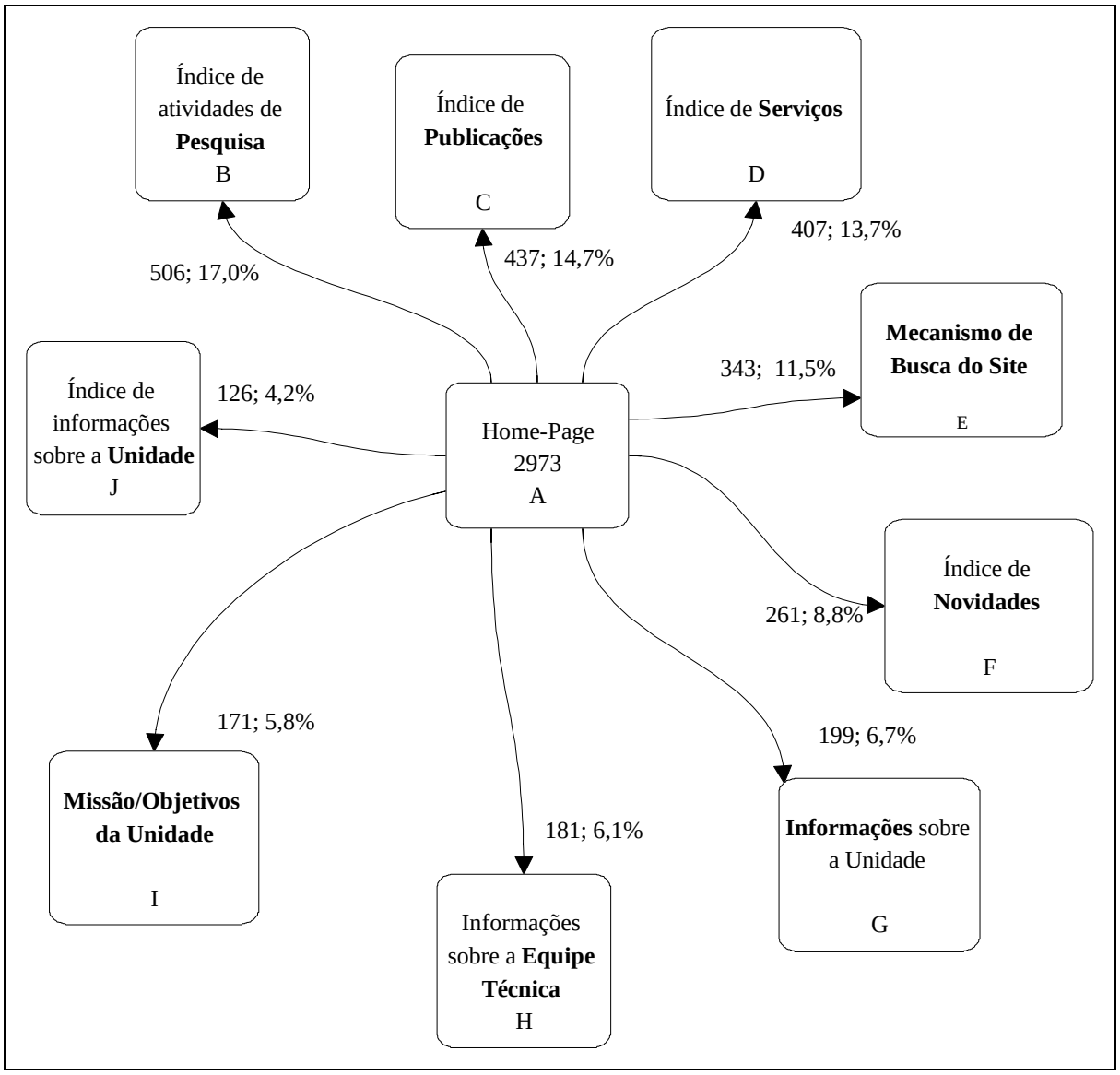

FIGURA 7 - Número de sessões por link acessado na primeira escolha do visitante do *site* (A-B*, A-C*, A-D*, A-E*, A-F*, A-G*, A-H*, A-I*, A-J*)

A Figura 8 apresenta o mesmo diagrama da Figura 7, não considerando, porém, a ordem de escolha dos conteúdos disponíveis na página principal. A ordem da freqüência dos *links* parece se manter quando as duas figuras são comparadas, com exceção do *link* "Mecanismo de Busca", que passa a ser o terceiro mais utilizado, posição ocupada pelo *link* "Serviços" quando considera-se a primeira escolha (Figura 7).

Já o *link* "Índice de Atividades de Pesquisa" é reafirmado como a página mais visitada do *site*, aparecendo em 3 de cada 10 sessões.

Foi observada uma relação muito forte entre as freqüências dos dois diagramas, retornando o coeficiente de correlação de 0,97 e ajustando-se perfeitamente a uma equação de reta (R-square = 0,95).

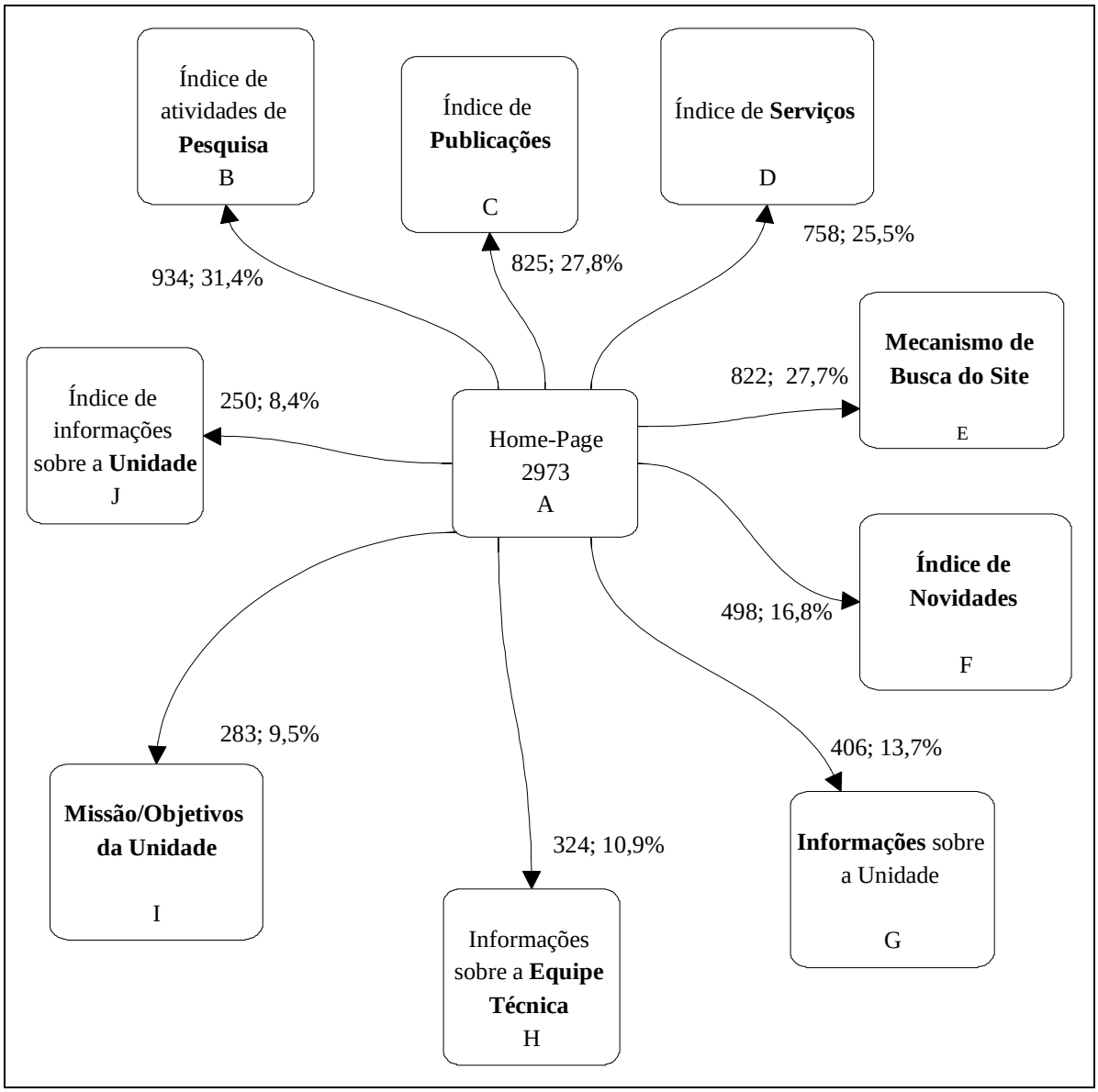

FIGURA 8 - Número de sessões nas quais foram acessados os principais *links* da *home-page* (A*B*, A*C*, A*D*, A*E*, A*F*, A*G*, A*H*, A*I*, A*J*)

Aprofundando a investigação, procurou-se saber quais eram as primeiras escolhas dos visitantes ao entrar nos três *links* da página principal, de maior preferência na primeira escolha ao entrar no *site*.

A Figura 9 mostra quais os caminhos tomados pelo visitante que escolhe o *link* "Pesquisa" na primeira escolha. O diagrama mostra que cerca de 3 em cada 10 visitantes (29,6%) encerravam a sessão após acessar o *link*, quantidade considerada alta. A segunda escolha de maior freqüência recaiu sobre o *link* "Projetos em Desenvolvimento" (15,8%), que apresentava, como conteúdo, uma listagem dos projetos e subprojetos de P&D conduzidos pela organização que abriga o *site*. Destes, 36,3% encerraram a sessão após o acesso e 15% passavam para a página contendo o "Índice de Publicações".

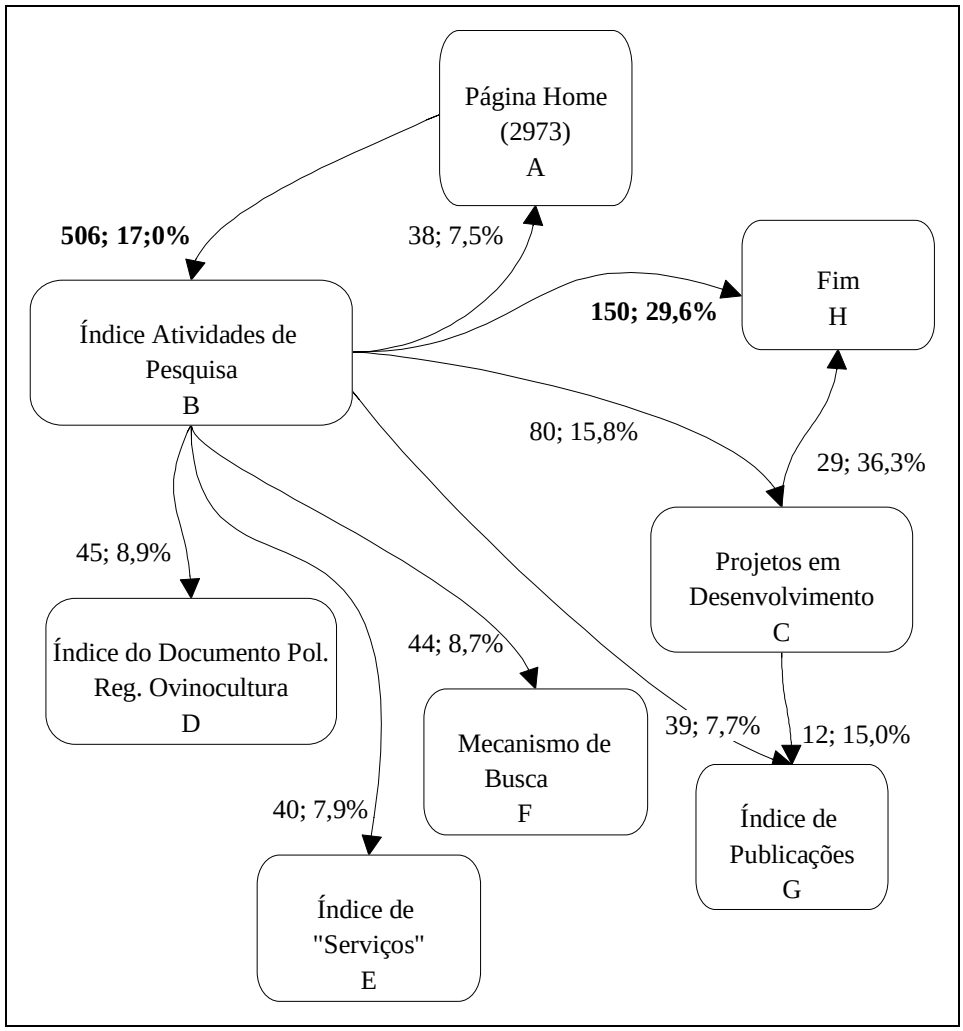

FIGURA 9 - Comportamento e preferências dos visitantes que acessaram o *link* "Pesquisa" como primeira escolha (A-B*, A-B-C*, A-B-C-G*, A-B-C-H, A-B-G*, A-B-D*, A-B-E*, A-B-F*, A-B-H, A-B-A*)

Já a Figura 10 reapresenta o diagrama da Figura 9 com mudanças de freqüências que merecem algumas considerações. O percentual de pessoas que continuaram a sair do *site* pela página “Índice de Atividade de Pesquisas” - e neste caso, após acessar outras páginas do *site*, também - não sofreu muita alteração (29,6% / 27,9%). O mesmo ocorreu com a saída pela página “Projetos em Desenvolvimento” (36,3% / 38,6%).

Entretanto, o percentual de acesso para os *links* “Mecanismo de Busca” e “Índice de Serviços” , assim como das sessões nas quais o visitante retornou à página principal, mais do que triplicou. Não há surpresa quanto ao primeiro, uma vez que o uso do mecanismo de busca como terceiro *click* aparece em apenas 5,8% das sessões que iniciam pela página principal, e 82,7% da freqüência deste recurso aparece do quarto *click* em diante. Já o aumento dos acessos ao *link* “Índice de Serviços”, bem como da volta à página principal, pode ter vários significados, merecendo uma investigação mais específica.

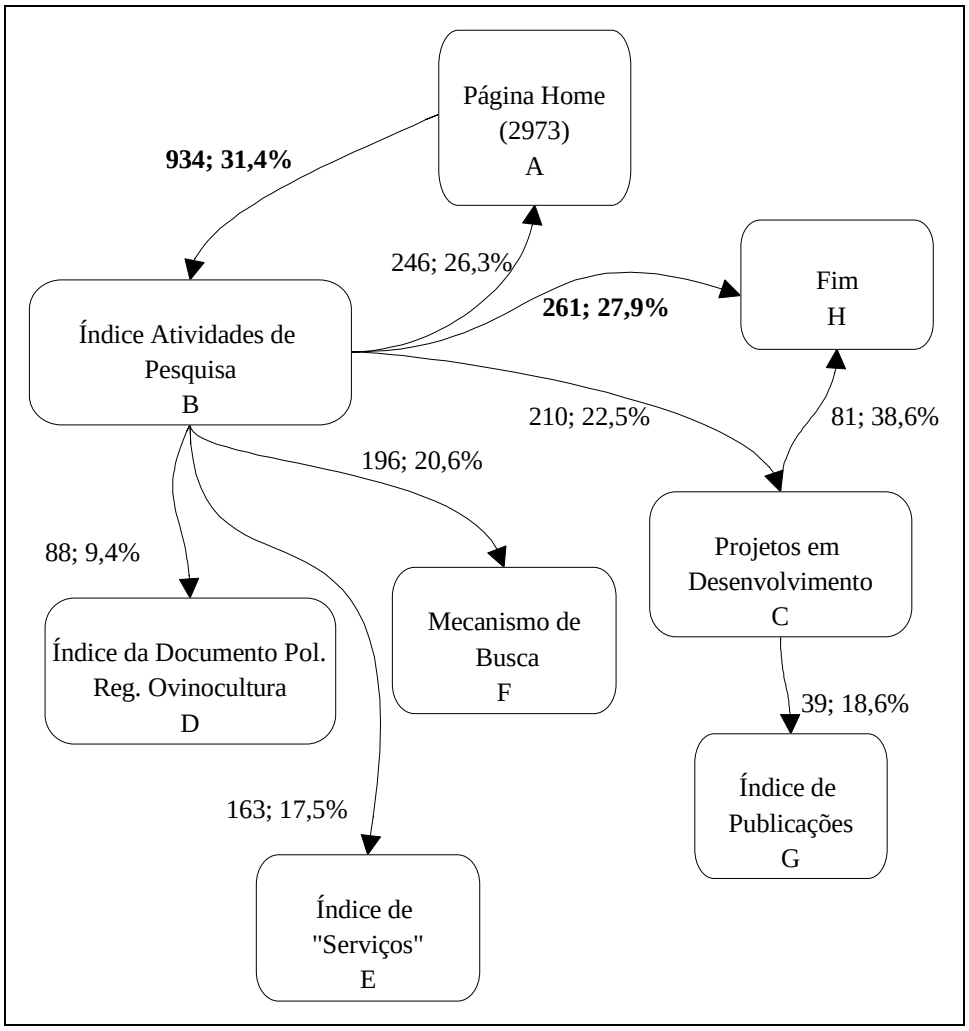

FIGURA 10 - Comportamento e preferências dos visitantes que acessaram o *link* "Pesquisa" durante a sessão (A*B*, A*B*C*, A*B*C*G*, A*B*C*H, A*B*G*, A*B*D*, A*B*E*, A*B*F*, A*B*H, A*B*A*)

O comportamento e as preferências dos visitantes que acessaram o *link* "Índice Documentos/Publicações" na primeira escolha pode ser visualizado na Figura 11. Pode-se ver que há bem menos "desistência" da sessão (10,1%) antes do segundo *click* do que os que acessaram o *link* "... Pesquisa" (Figuras 9 e 10). O interesse principal recai sobre o "Catálogo de Publicações" (34,6%). Todavia, 32,5% dos visitantes que acessaram este *link* - por este caminho - encerraram a sessão nele, sendo também baixa a volta à página principal.

Após o acesso ao "Catálogo de Publicações", a primeira escolha foi o "Índice da Folha do Produtor" (15,9%) , veículo de versão *on-line* que publica artigos, entrevistas e relatos das atividades de P&D da unidade de pesquisa. Este *link* também foi acessado diretamente a partir

do “Índice Documentos/Publicações” em 21,1% das sessões.

Um aspecto interessante sobre este *link* é que os visitantes que preencheram o formulário *on-line* para receber o informativo na versão impressa, tiveram dados muito acima da média: tempo da sessão 1443,5s e número de *pageviews* 12,2. O mesmo aconteceu com as sessões nas quais os visitantes acessaram o documento sobre políticas para ovinocultura (1124,2 e 12,5);

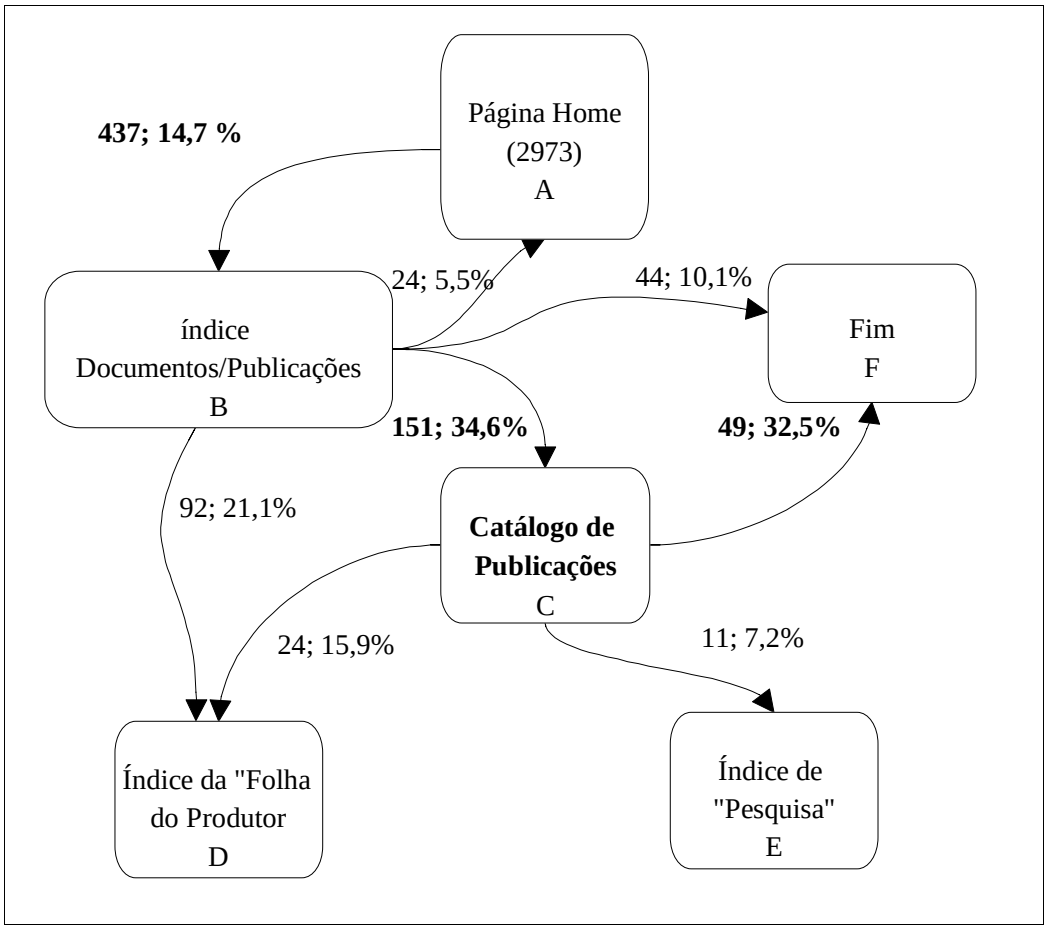

FIGURA 11 - Comportamento e preferências dos visitantes que acessaram o *link* “Publicações” na primeira escolha (A-B*, A-B-C*, A-B-C-D*, A-B-C-E*, A-B-C-F*, A-B-D*, A-B-F*, A-B-A*)

A Figura 12 não mostra grandes alterações em relação a Figura 11. O percentual de desistência da sessão em B e em C praticamente se mantém (12,2%).

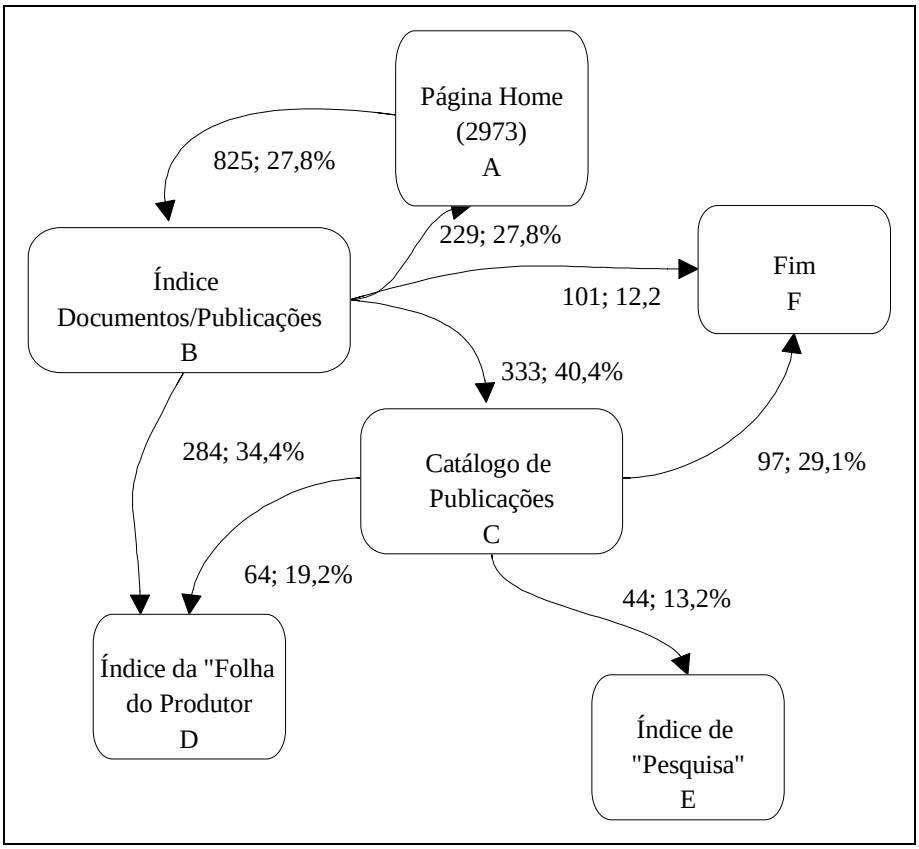

FIGURA 12 - Comportamento e preferências dos visitantes que acessaram o *link* "Publicações" durante a sessão (A*B*, A*B*C*, A*B*C*D*, A*B*C*E*, A*B*C*F, A*B*D*, A*B*F, A*B*A*)

Nas Figuras 13 e 14 são apresentadas o comportamento e as preferências dos visitantes que tiveram com o primeiro *click* o "Índice de Serviços". Este *link*, o qual agrega alguns dos serviços, tecnologias e produtos disponibilizados pela organização, se reveste de especial importância, uma vez que representa o que a empresa tem a oferecer à sociedade como forma de retorno ao financiamento das suas atividades.

A Figura 13, em particular, mostra a primeira decisão dos visitantes cujo primeiro *click* é o *link* "Índice de Serviços". Nota-se que as taxas de desistência da sessão e de retorno à página principal não são muito diferentes das taxas dos diagramas já apresentados anteriormente (23,8% e 6,8%, respectivamente). Após acessar o *link*, a primeira escolha recai sobre o

portfólio de tecnologias, serviços e produtos gerados pela unidade de pesquisa e colocados a disposição de seus usuários e clientes (42,8%). Destes, 16,7% desistem da sessão neste *link*.

As próximas escolhas ficam, então, pulverizadas, recaindo a escolha mais significativa sobre o *link* "Consultoria em Nutrição Animal" (15,5%). Seguindo esta opção, o *link* "Laboratório de Nutrição Animal" apresenta-se como a segunda escolha de maior freqüência (10,9%). Como as duas estão intimamente relacionadas e, juntas representam um percentual de 26,4% (1 a cada 4 visitantes), estes dados mereceriam maiores considerações pelos mantenedores do *site*, podendo indicar uma demanda, uma oportunidade de negócios para a empresa ou a necessidade de qualificar os serviços e os conteúdos disponibilizados.

Já a procura pelas espécies forrageiras (*links* D e G) não foi tão significativa, somando juntas 13,2%, sendo a de maior preferência dos visitantes a espécie *trifolium Repens* (Trevo Branco).

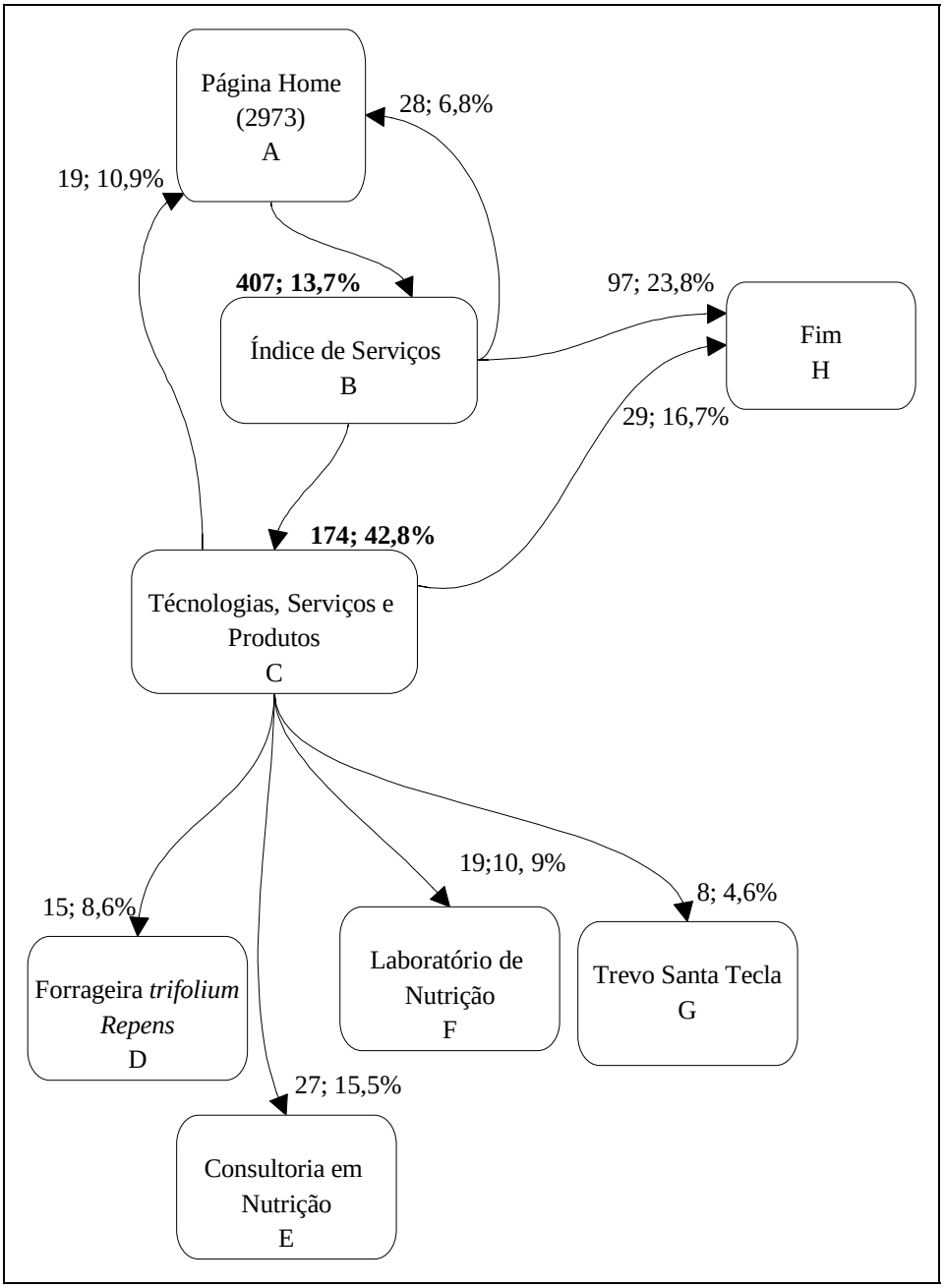

FIGURA 13- Comportamento e preferências dos visitantes que acessaram o *link* "Serviços" na primeira escolha (A-B*, A-B-A*, A-B-C-A*, A-B-C-D*, A-B-C-H, A-B-C-E*, A-B-C-F*, A-B-C-G*, A-B-C-H)

A Figura 14 apresenta o mesmo diagrama apresentado na Figura 13, porém com duas adições (*links* H e I). Estes *links* tiveram freqüência muito baixa quando foi considerada apenas a primeira escolha do visitante e, desta forma, não foram incluídos no diagrama da Figura 13.

Os percentuais dos dois diagramas não apresentam muitas modificações - em termos ordinais

- quando analisamos os *links* que apontam para as tecnologias, serviços e produtos, tendo o conteúdo sobre “Consultoria em Nutrição” confirmado a preferência dos visitantes (25,0%) e somando, juntamente com o *link* “Laboratório de Nutrição”, 36,5% de freqüência na preferência dos usuários que chegam à página do portfólio.

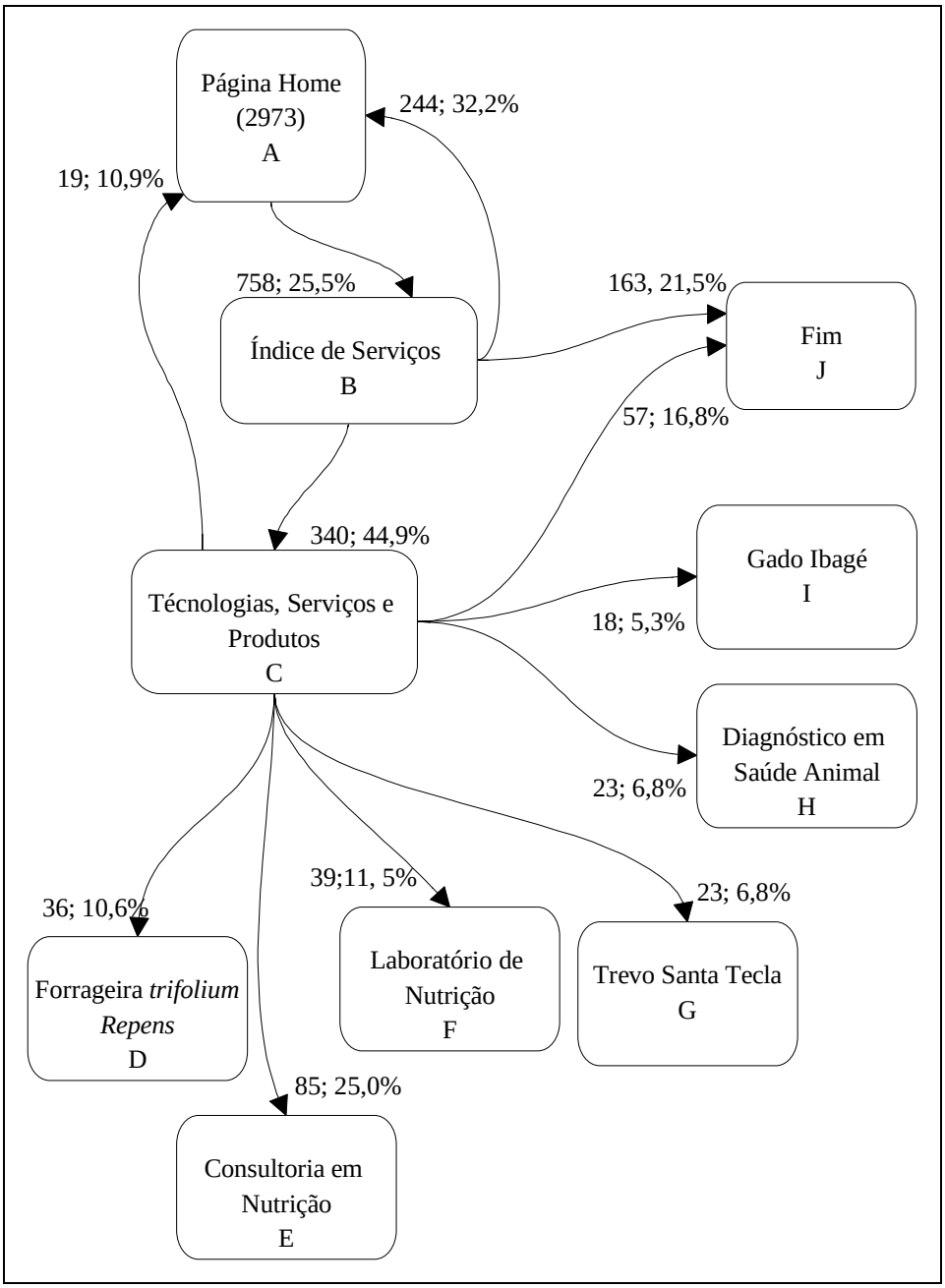

FIGURA 14 - Comportamento e preferências dos visitantes que acessaram o *link* “Serviços” durante a sessão (A*B*, A*B*A*, A*B*C*A*, A*B*C*D*, A*B*C*H, A*B*C*E*, A*B*C*F*, A*B*C*G*, A*B*C*H*, A*B*C*I*, A*B*C*J, A*B*J)

4.3 Análise e discussão dos termos de consulta utilizados pelos usuários do *site*

Foram coletadas, analisadas e classificadas palavras-chave inseridas no mecanismo de busca do *site* entre às 18h25 horas de 08/11/1999 e 11h09 horas de 15/06/2000. A amostra totalizou 2.905 termos para busca coletados em 1.473 sessões. Neste conjunto de dados não foi aplicado nenhum critério de seleção, sendo consideradas todas as consultas realizadas no *site*.

Em média, o tempo entre consultas na mesma sessão ficou 169,6 seg. Já o tempo total em que o visitante permaneceu no *site* ficou em 327,4 seg. Das 2.905 consultas, 1.191 não tiveram o endereço resolvido, não sendo possível identificar a origem. As restantes (1.714) estavam distribuídas conforme a Tabela 14.

Observa-se que a distribuição das consultas pelos diferentes domínios mostra pequenas variações em relação a distribuição das sessões apresentada na Tabela 4, a qual registra uma queda de 6,0 pontos percentuais na freqüência das instituições de ensino e pesquisa, e uma elevação de 4,8 pontos no uso do mecanismo pelo domínio comercial e ISPs.

TABELA 14 - Distribuição das consultas considerando a origem do visitante

Número de consultas	**Freqüência**	**(%)**
Comercial/ISPs	1.314	76,7
Inesituições de Ensino e de Pesquisa	252	14,7
Instituições de Ensino e de Pesquisa do RS	26	1,5
Provedores de Rede	70	4,1
Governo	17	1,0
Outros	35	2,0
Total	1.714	100,0

A fim de ter uma idéia sobre o conteúdo das consultas, procurou-se classificar cada uma considerando sua pertinência ao contexto ao contexto do *site*. As classes utilizadas foram:

Dentro do contexto da unidade de pesquisa - termos diretamente relacionados à missão da organização. Ex.: bovino, ovino, pastagem, pecuária, leite, carrapato, etc.

Fora do contexto da unidade mas dentro do contexto da Embrapa - termos não cobertos pela missão da unidade, mas cobertos pela missão da Embrapa. Ex.: suíno, aves, psicultura, bubalinos, pêssego, etc.

Fora do contexto da unidade e da Embrapa, mas dentro do contexto do agronegócio - termos não cobertos pelas missões da unidade e/ou da Embrapa ou para os quais a unidade e/ou Embrapa não se apresentam como referência. Ex.: scargot, cunicultura, ranicultura, chichila, codorna, etc.

Fora do contexto do agronegócio - termos que não apresentavam relação direta com o agronegócio. Ex.: carvão mineral, sabão de abacate, vinagrete, ensaio de dureza, etc.

Dúbio ou não identificado - termos muito genéricos que deixaram dúvidas quanto à sua classificação. Ex.: custo de produção (de quê ?), dados econômicos (sobre o quê ?), folhas (

de quê ?), etc.

Informações administrativas - termos de consulta utilizados para acessar informações de caráter mais administrativo da unidade ou da Embrapa. Ex.: estágio, concursos, leilão, e-mails, endereço, biblioteca, etc.

Busca por receitas - termos utilizados para buscar receitas de preparação de carne do *site*. Ex.: pernil de cordeiro, receitas, churrasco, carnes, receitas com cordeiro, etc.

Os resultados são apresentados na Tabela 15. Nota-se que o percentual de consultas relacionadas diretamente à missão da Unidade é 64,2%, totalizando, junto com as consultas cobertas pela missão da Embrapa, 80,1%. Este percentual, junto com o percentual das consultas, fora das missões da unidade e Embrapa, mas dentro do contexto do agronegócio, totalizou 83,7% de pertinência.

Um pequeno percentual de termos foi considerado de sentido dúbio ou não identificado (6,1%), o que pode apontar, entretanto, para a necessidade de desenvolvimento de assistentes de navegação dotados de mais inteligência.

O percentual de 3,2% de consultas classificadas como fora do contexto do agronegócio pode ser considerado ínfimo, demonstrando ter o visitante um foco bem definido.

TABELA 15 - Classificação dos termos para pesquisa utilizados no mecanismo de busca do *site*

Contexto da consulta	**Freqüência**	**(%)**
Dentro do Contexto da Unidade de P&D	1.866	64,2
Fora do contexto da Unidade mas dentro do contexto da Embrapa	463	15,9
Fora do contexto da Unidade e da Embrapa, mas dentro do contexto do agronegócio	105	3,6
Dúbio ou não identificado	177	6,1
Informações administrativas	128	4,4
Fora do contexto do Agronegócio	94	3,2
Busca por Receitas	72	2,5
Total	2.905	100,0

A Tabela 16 exibe a freqüência de consultas por sessão. Nota-se que aproximadamente metade dos visitantes (50,7%) realizou apenas 1 consulta, e que grande parte realizou até 3 consultas no *site* (87,8%). Também observa-se que a cada incremento de 1 no número de consultas por sessão, a freqüência da classe cai em média 50%.

TABELA 16 - Número de consultas realizadas por sessão

Número de consultas por sessão	Freqüência	(%)
1	1.473	50,7
2	741	25,5
3	336	11,6
4	171	5,9
5	87	2,9
6	41	1,4
7	23	0,8
8	16	0,6
9	6	0,2
>9	11	0,4
Total	2.905	100,0

Os termos dentro do contexto da unidade, da Embrapa e do agronegócio, totalizaram 2.426. Estes foram sumarizados em 1.248 termos diferentes. Os termos utilizados 10 vezes ou mais, na exata forma como foram digitados e considerando aqueles contextos, estão apresentados na Tabela 17. Observa-se que com exceção dos termos “caprinos”, “suínos” e “suinocultura”, os restantes estão relacionados diretamente com a missão da unidade. Todos os termos apresentados estão cobertos pela missão da Embrapa. Ressalta-se que os termos apresentados na Tabela 17 - apenas 1,3% dos termos utilizados - apareceram em 20,4% das consultas. Registra-se o termo “história da pecuária”, o qual pode indicar uma oportunidade de promoção.

TABELA 17 - Freqüência dos termos mais utilizados no mecanismo de busca do *site*, da forma em que foi digitado pelo visitante

Termo utilizado pelo visitante	Freqüência
ovinos	81
pecuária	67
ovinocultura	58
confinamento	48
campos	37
pastagens	29
gado de corte	28
pastagem	27
ovino	17
leite	16
bovinos	15
caprinos	14
gado	14
suínos	13
gado de leite	11
história da pecuária	10
suinocultura	10
Sub-total	495
(%)	(20,4)
outros	1.931
(%)	(79,6)
Total	2.426
(%)	(100,0)

Notas: registre-se uma freqüência de 31 sessões para o termo “receitas”, o qual não aparece nesta tabela por ter sido classificado separamente. O termo “nutrição” aparece também em 11 sessões. Entretanto, foi classificado como “dúbio” e não aparece na tabela.

A fim de agregar um pouco mais os termos utilizados pelos visitantes, procurou-se classificá-los segundo o Thesagro - um manual de catalogação de documentos elaborado pelo Ministério da Agricultura e do Abastecimento e utilizado por documentalistas da área agrícola. Dos 2.434 termos dentro do contexto da unidade, da Embrapa e do agronegócio, 72 não foram classificados por aquele sistema, sobrando, então, 2.362 termos que foram reduzidos para 435 termos diferentes após a sumarização. Aqueles que apresentaram freqüência igual a 20 ou mais estão apresentados na Tabela 18.

Nota-se, novamente, grande aderência dos termos utilizados com a missão da unidade de pesquisa, apresentando somente uma exceção (“caprinos”). Constata-se que 42,5% das consultas giravam em torno de 4,1% dos termos. Novamente, todos os termos apresentados estão cobertos pela missão da Embrapa. Ressalta-se a confirmação do interesse pelos assuntos

“ovino” e “ovinocultura”, assim como assuntos relacionados com a nutrição de rebanhos (“confinamento”, “pastagem”, “capim”, “planta forrageira”, “nutrição animal” e “campo”).

TABELA 18 Freqüência dos termos mais utilizados no mecanismo de busca do *site*, após classificação pelo Thesagro

Termo classificado pelo Thesagro	**Freqüência**
pecuária	188
ovino	155
confinamento	89
pastagem	81
ovinocultura	64
campo	62
gado de corte	53
instalação para animal	46
nutrição animal	32
leite	30
bovino	29
gado leiteiro	28
gado	27
ovelha	27
doenca animal	26
planta forrageira	23
caprino	22
capim	22
Sub-total	1.004
(%)	(42,5)
outros termos	1.358
(%)	(57,5)
Total	2.362
(%)	(100,0)

Procurando-se averiguar o interesse dos visitantes focando apenas as espécies animais implícitas ou explícitas na sessão, classificou-se as consultas e calculou-se a freqüência conforme as atividades apresentadas na Tabela 19. Desta forma pode-se sintetizar mais ainda as consultas.

Observa-se que a atividade “Bovinocultura” possui a maior freqüência, em contraposição aos dados já apresentados nas Tabelas 17 e 18, que demonstravam uma maior freqüência de termos relacionados à “Ovinocultura”. Deve-se isto a vários fatores, entre os quais o fato de o termo aglutinar tanto bovinos de leite quanto de carne, o fato de que o usuário que procurou por “Bovinocultura” pareceu ser mais específico em suas consultas e o fato de o Thesagro ter

mais divisões para aquela atividade do que para a atividade de “Ovinocultura” (chegando ao nível de raças, p.ex.).

Nota-se também que, 82,7% das consultas buscavam as espécies animais cobertas pela missão da Unidade (bovinocultura e ovinocultura), devendo-se considerar, ainda, o percentual de 10,6% que buscavam por suinocultura, caprinocultura e avicultura, termos estes fora do contexto da missão da unidade de pesquisa, mas dentro da missão da Embrapa. Um aspecto interessante da Tabela 19, é o fato da relação explícita da consulta e implícita na sessão ser muito maior naquelas sessões que buscavam bovinocultura.

TABELA 19 - Classificação das consultas segundo a espécie animal implícita na consulta e explícita na sessão

Atividade de criação	**Explícita na consulta**	**Implícita na sessão**
Bovinocultura	412	161
Ovinocultura	377	21
Caprinocultura	43	1
Suínocultura	42	6
Avicultura	32	1
Psicultura	23	0
Bubalinocultura	13	1
Eqüinocultura	14	0
Outras atividades	27	0
Totais	983	191
Total Geral		1.174

Nota: 24 consultas envolviam mais do que uma espécie e 191 estavam interessados em pecuária como um todo.

Através da discussão apresentada, procurou-se enfocar o objeto de estudo (*site*) sob diferentes aspectos. Através da quantificação das variáveis básicas de acesso ao *site*, da determinação das preferências primárias de navegação e da análise das necessidades explícitas pelos usuários no mecanismo de busca, pode-se formar um perfil da demanda de informações pelos visitantes. No capítulo seguinte procurar-se-á sintetizar os resultados obtidos através da

convergência dos diferentes aspectos abordados.

CAPÍTULO 5: Conclusões

Considerando os referenciais teóricos do método adotado nesta pesquisa - estudo de caso de nível exploratório – tornam-se necessárias algumas considerações, não somente a respeito dos resultados obtidos, mas também, sobre alguns aspectos práticos e metodológicos a serem considerados em estudos desta natureza, além dos já mencionados no item 3.3.1. Finalmente, extrapolando o caráter técnico da investigação, resumidamente serão feitas algumas considerações sobre os seus potenciais benefícios, considerando uma perspectiva socio-econômica mais ampla.

5.1 Quanto ao objetivo do estudo

As análises efetuadas elucidaram alguns aspectos antes obscuros quanto à audiência do *site* estudado. Além de aspectos relacionados à acessibilidade e freqüência, considerando diferentes distribuições no tempo, foi possível também formar uma idéia sobre o comportamento de navegação do visitante, bem como sobre suas necessidades e preferências explícitas por informações ligadas à pecuária.

A obtenção de métricas sobre as características das transferências de conteúdo, freqüência das sessões e tempos de exposição poderão ser úteis para o planejamento dos recursos físicos do *site*. Estas métricas poderão ser consideradas em estimativas de acessos, ou para objetivos mais específicos como balanceamento de carga (*load-balance*) em servidores.

A determinação das origens das requisições pode fornecer subsídios para o fortalecimento de relações com instituições congêneres e potenciais clientes e usuários, podendo ser utilizada pelas atividades de comunicação e de marketing, bem como consideradas quando na determinação de ações estratégicas da organização, principalmente aquelas voltadas à articulação com seu ecossistema.

Quanto às preferências de navegação, a primeira escolha dos visitantes girou basicamente em

torno da necessidade de informações sobre as atividades de pesquisa desenvolvidas, as publicações produzidas e os serviços oferecidos pela instituição. Já os termos utilizados no mecanismo de busca, giraram predominantemente em torno das atividades de bovinocultura e ovinocultura e seus aspectos relacionados.

Nota-se, assim, que, de uma maneira geral, os aspectos estudados indicaram que a demanda por informações no *site* apresenta aderência com a missão da organização que o mantém. As expectativas explícitas dos visitantes pouco desviaram do conteúdo disponibilizado e propagado pelo *site* estudado. Reconhece-se, porém, a ausência de métodos capazes de medir de forma padronizada, o alinhamento entre as características de acesso ao *site* e missão da organização que o mantém.

As exceções observadas, entretanto, merecem considerações por parte da organização, podendo indicar ações que visem aumentar o nível de atendimento das necessidades do visitante, através do redirecionamento para *sites* congêneres que possam atende-lo, principalmente aqueles localizados em outras unidades de pesquisa da mesma organização. Podem também mostrar que a missão da organização deve ser melhor propagada, diminuindo assim o nível da demanda por informações não pertinentes ao escopo do *site*.

O estudo efetuado poderá contribuir para a elaboração de assistentes virtuais que possam direcionar os visitantes para os temas procurados, como auxílio inteligente à transferência de informações e de inovações tecnológicas, podendo ser também considerado na determinação dos *links* e conteúdos com mais foco.

5.2 Quanto aos aspectos práticos e metodológicos

Embora existam ferramentas desenvolvidas com o objetivo de apoiar a investigação da audiência de *sites* através da análise de *logs* gerados pelas aplicações, sua utilização ainda não é amplamente disseminada. Menos disseminadas ainda são aquelas ferramentas, para o mesmo fim, mas que incorporam técnicas de mineração de dados e facilidades de consulta livre pelo analista.

Geralmente, dados de acessos aos *sites*, por ocuparem considerável espaço em disco, são

desprezados e considerados como um “incômodo” pelos administradores de sistemas, não sendo difícil encontrar *sites* em que o *log* de transações esteja desabilitado.

Todavia, a obtenção destes dados, mesmo quando facilitada, não se traduz em certeza de um conjunto de dados pronto para ser analisado. O desenho do *site* e sua concepção navegacional são fatores que devem, quando possíveis, preceder ao trabalho de investigação, com vistas a facilitar tanto a preparação, quanto a análise dos dados. É necessário que seja feita esta ressalva, uma vez que, o *site* analisado já existia antes da investigação, tendo, este fator, dificultado a análise dos dados, principalmente quando estes foram submetidos ao *software* de geração de regras C4.5.

Uma das etapas críticas é o processo de preparação dos dados, o qual envolve aspectos independentes e aspectos dependentes do próprio *site.* Entre os aspectos independentes está a limpeza do arquivo, que vai depender muito do objetivo do estudo. Em geral, elementos gráficos, acessos de *robots*/*spiders*, requisições interrompidas ou não encontradas, etc não são mantidos no conjunto final de dados.

Entre os processos dependentes do *site*, está a técnica de subdividir as sessões em transações (COOLEY *et. al*., 1997b), extremamente útil e necessária no sentido de tornar menos complexa a análise, principalmente quando forem utilizadas técnicas de mineração de dados e/ou quando o número de páginas do *site* é muito grande. A distinção entre páginas de navegação e páginas de conteúdo também pode ser interessante para a redução do arquivo a ser utilizado, bem como para a redução da complexidade da análise.

De uma maneira prática, deve-se buscar simplicidade e objetividade na organização dos conteúdos e seus *links*, com vistas a ter-se uma idéia clara do que representa cada c*lick* - ou conjunto de *clicks* dado pelo visitante (*clickstream*). Logicamente, este componente pressupõe conhecimentos específicos do investigador em relação ao *site* analisado.

Por serem configurados para permitir acesso somente por login/password, os ambientes de ensino e treinamento virtuais ou à distância, em particular, se apresentam como um campo promissor para pesquisas desta natureza. Todavia, é necessário o desenvolvimento de ferramentas capazes de automatizar o processo de análise de *logs*. Devido ao amplo leque de

questionamentos possíveis, estas ferramentas devem ser concebidas de maneira a agregar uma linguagem de consulta livre para o analista.

O tema tem potencial para suscitar investigações de níveis descritivo, exploratório e explicativo. Este último, em particular, pode ser considerado quando da execução de experimentos em que o monitoramento do uso de sistemas remotos acontece no lado do cliente (*client-side*), p.ex. (FREITAS, 1993; SAKAMOTO, 1997). Já o estudo de caso como estratégia de pesquisa, parece ser a abordagem mais adequada, dada as particularidades inerentes à atividade-fim da empresa, sua inserção no ecossitema como organização, a sua história e a do *site*, aspectos relacionados ao *design* e tantos outros aspectos que o tornam "único". Característica esta, aliás, perseguida na etapa de concepção dos ambientes virtuais.

5.3 Contribuições potenciais

A Embrapa Pecuária Sul é uma unidade de pesquisa e desenvolvimento (P&D) descentralizada da Empresa Brasileira de Pesquisa Agropecuária - Embrapa, localizada em Bagé-RS, que tem a missão institucional de "Atender com soluções tecnológicas as necessidades dos sistemas produtivos integrados ao agronegócio de bovinos e ovinos na Região Sul em benefício da sociedade".

Seus usuários/clientes podem ser caraterizados por produtores rurais mais ligados à pecuária, estudantes de agronomia, veterinária e de escolas técnicas de agropecuária, sindicatos rurais, associações de produtores, extensionistas rurais, cooperativas, órgãos de fomento, outras instituições de pesquisa agropecuária, universidades e empresas públicas ou privadas envolvidas com as cadeias produtivas da carne (ovina e bovina) e do leite. Sua região de abrangência é entendida como sendo os estados da Região Sul (PR, SC, e RS).

Uma agropecuária competitiva pressupõe uma forte atividade de pesquisa agropecuária. Esta atividade, entretanto, deve possuir mecanismos eficientes e eficazes de transferência da tecnologia gerada. Um destes mecanismos poderá ser a *Web*.

Assim, o benefício que os resultados deste estudo podem trazer para a sociedade está apoiado na premissa de que uma maior e mais direcionada oferta de informações tecnológicas pode contribuir para a aceleração do processo de geração/aplicação de conhecimentos advindos da

pesquisa tecnológica agropecuária. Esta aceleração, por sua vez, poderá contribuir para o aumento dos índices de produtividade agropecuária impactando toda a cadeia produtiva da carne e do leite, com reflexos positivos na renda do produtor rural e da cadeia como um todo.

Neste caso, saber o que os usuários - atuais e potenciais - de informação em tecnologia de produção agropecuária desejam, ou desejariam, encontrar sobre tecnologias agropecuárias na *Web* poderá ter grande utilidade. Isto, além de possibilitar o alinhamento das estratégias da organização as necessidades da matriz produtiva, possibilita também que a organização identifique e monitore áreas de crescente interesse que possam ser utilizadas para impactar a eficiência e eficácia na transferência de seus resultados. As informações disponibilizadas poderão, em algum nível, contribuir para o processo de tomada de decisões que permeia todos os segmentos da cadeia produtiva dos produtos considerados.

Já para a área de SI, a análise do ecossistema virtual da organização poderá apontar recomendações sobre a forma e o conteúdo das páginas do *site*, de forma a torna-lo adaptativo aos usuários. A pesquisa poderá contribuir, também, para estudos que visem alinhar esta TI emergente (*Web*) com as estratégias e objetivos de uma organização, no sentido de que pode apontar oportunidades de negócios, de parcerias, áreas de interesse crescente, etc.

Além de recomendações, o projeto poderá contribuir para a especificação ou melhoria de uma ferramenta - ou metodologia - de análise de *logs*, que monitore acessos ao *site* a partir de um enfoque de maior utilidade para a área de *marketing* estratégico. Poderá, também, apontar a necessidade de bases de dados estruturadas, modeladas a partir das necessidades dos clientes e usuários que acessam o *site*.

Para o aluno, a identificação de elos entre a sua prática diária e o corpo teórico técnico-científico da área de administração de SI contribui, por um lado, para o alargamento da sua base teórica, e de outro, para aumentar suas qualificações como profissional da área de TI. Uma melhor compreensão de métodos de pesquisa quantitativa e qualitativa também é um fator que contribui para o aperfeiçoamento de seus métodos e técnicas no ensino universitário.

Em relação ao contexto imediato, as recomendações do estudo poderão indicar uma reconfiguração das páginas do *site* no sentido de torná-lo alinhado com as expectativas dos

usuários, clientes, parceiros e visitantes ocasionais. Apontar possíveis oportunidades estratégicas e nichos de consumo para informações tecnológicas no âmbito da missão da unidade considerada são, também, possibilidades.

Os resultados poderão ajudar também na identificação de oportunidades para promoção de negócios tecnológicos, na identificação de oportunidades para a promoção da imagem da empresa, na transferência de tecnologias e divulgação das atividades da empresa junto ao seu mercado-alvo, em futuros estudos visando implementar transações comerciais eletrônicas entre a Embrapa, e seus clientes, e parceiros e na elaboração de *FAQs* (*Frenquently Asked Questions*) sobre assuntos de interesse de clientes, parceiros e visitantes ocasionais.

REFERÊNCIAS BIBLIOGRÁFICAS:

ABDULLA, G.; FOX, E.A. e ABRAMS, M. Shared User Behavior on the World Wide Web. Proceedings of the WebNet 97 – World Conference of WWW, Internet and Intranet. Toronto-Canada, Nov., 1-5, 1997.

ARLITT, M.F. e WILLIAMSON, C.L. Web server workload characterization: the search for invariants. Proc. SIGMETRICS, Philadelphia, PA, April 1996. ACM, 160-169.

BAMSHAD, M.; COOLEY, R. SRIVASTAVA; J. Automatic Personalization Based on Web Usage Mining... Available from World Wide Web: <http://http://maya.cs.depaul.edu/~mobacher/personalization/.>, consulta em mar./2000.

BERNES-LEE, T. Frequently Asked Questions. Available from World Wide Web: <http://www.w3.org/People/Bernes-Lee/FAQ.html>, consulta em abr./1999.

BÜCHNER, A.G.; ANAND, S.S.; MULVENNA, M.D. e HUGHES, J.G. Discovering Internet Marketing Intelligence Trough Web Log Mining. Available from World Wide Web: <http://www.infj.ulst.ac.uk/~cbgv24/ papers/ Unicom99.pdf>, consulta em mar./2000.

CASTELLS, M. A sociedade em rede. São Paulo: Paz e Terra, 1999. 617p

CHEN, M.S.; PARK, J.S. e YU, P.S. Data Mining for path traversal patterns in web environment. In Proceedings of 16th International Conference on Distributed Computing Systems, 1996.

CSIKSZENTMIHALYI, M. Beyond boredom and anxiety, second printing. São Francisco: Jossey-Bass, 1977.

CSIKSZENTMIHALYI M. e LEFEBRE J. Optimal experience in work and leisure. Journal of personality and social psychology, 56 (5), 815-822, 1989.

CSIKSZENTMIHALYI, M. Flow: The psycology of optimal experience. New York: Harper and Row, 1990.

COOLEY, R.; MONBACHER, B. e SRIVASTAVA, J. Web Mining: Information and pattern discovery on the world wide web. Proceedings of *ICTAI'97*. Newport Beach, California. 3-8 Nov., 3-8, 1997a. Available from World Wide Web: <http://maya.cs.depaul.edu/~mobasher/webminer/survey/survey.html>, consulta em mar./2000.

COOLEY, R.; MONBACHER, B. e SRIVASTAVA, J. Grouping Web References into Transactions for Mining World Wide Web Browsing Patterns. Technical Report TR 97-027, University of Minessota, Dept of Computer Science, Minneapolis, 1997b. Available from World Wide Web: <http://maya.cs.depaul.edu/ ~mobasher/pubs-subject.html>, consulta em mar./2000.

COOLEY, R.; MONBACHER, B. e SRIVASTAVA, J. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems. 1 (1998) 00-00. Available from World Wide Web: <http://maya.cs.depaul.edu/ ~mobasher/pubs-subject.html>, consulta em mar./2000.

ESTRATÉGIA GERENCIAL DA EMBRAPA: GESTÃO 95/98. Empresa Brasileira de Pesquisa Agropecuária. D.E., Brasília, 1995. 27p.

ESTRATÉGIA GERENCIAL DA EMBRAPA: MACROPRIORIDADES/1997. Empresa Brasileira de Pesquisa agropecuária. D.E., Brasília, 1997, 27p.

FLORES, M.X. Projeto EMBRAPA: a pesquisa agropecuária rumo ao século XXI. Brasília: EMBRAPA-SEA, 1991. 38p. (EMBRAPA-SEA. Documentos, 4).

FLORES, M.X. e SILVA, J. de S. Projeto EMBRAPA II: do projeto de pesquisa ao desenvolvimento sócio-econômico no contexto do mercado. Brasília: EMBRAPA-SEA, 1992. 55p. (EMBRAPA-SEA. Documentos, 8).

FREITAS, H. A Informação como ferramenta gerencial. Porto Alegre: Ortiz, 1993. 355p.

GATES, B. A empresa na velocidade do pensamento. São Paulo: Companhia das Letras, 1999. 444p.

GIL, A.C. Como elaborar projetos de pesquisa. 3.ed. São Paulo: Atlas, 1996. 159p.

GIL, A.C. Métodos e técnicas de pesquisa social. 5. ed. São Paulo: Atlas, 1999. 206p.

HOFFMAN, D.L. e NOVAK, T.P. Marketing in Hypermidia Computer-Mediated-Environments: Conceptual Foundations. Journal of Marketing, july, 1996. Available from World Wide Web: < http://www2000.ogsm.vanderbilt.edu/cme.conceptual. foundations.html >, consulta em mar./2000.

GONÇALVES, C. A. e FILHO, C. G. Tecnologia da Informação e *Marketing*. Como obter clientes e mercados. Revista de Administração de Empresas – RAE. São Paulo, v.35, n.4, p.21-32. Jul./Ago, 1995.

JANSEN, B.J.; SPINK, A.; BATERMAN, J. e SARACEVIC, T. Searchers, the subject they search, and sufficiency: a study of a large sample of Excite searchers. Proceedings of the WebNet 98 – World Conference of WWW, Internet and Intranet. Orlando-Flórida, Nov., 1-12, 1998.

JOHN, G., PANAGIOTIS, M.D. How to use HTML page popularity to improve a web site´s structure. Proceedings of the WebNet 98 – World Conference of WWW, Internet and Intranet. Orlando-Flórida, Nov., 1-12, 1998.

KOTLER, P. Administração de Marketing; análise, planejamento, implementação e controle. 8. ed. São Paulo: Atlas, 1996. ?p.

MONTGOMERY, D. B. e WEINBERG, C. B. Toward Strategic Intelligence Systems. Marketing Management, Winter, 1998.

NOVAK P.T e HOFFMAN, D.L. New metrics for new media: toward the development of Web measurement standards. Vanderbilt University, 1996. Available from World Wide Web: <http://www2000.ogsm.vanderlbilt.edu /novak/web.standards/ webstand.html, consulta em fev./2000.

LIVRO VERDE DA SOCIEDADE DA INFORMAÇÃO NO BRASIL. Grupo de Implantação do Programa Sociedade da Informacão. Ministério da Ciência e Tecnologia / MCT. Brasília: SocInfo, 2000. 90p. Disponível na World Wide Web: <http://www.socinfo.org.br>

OHMAE, K. Começando de novo. Revista HSM Management, n.11, ano 2, nov/dez 1998. p.6.

PENZIAS, A. Lições de um Prêmio Nobel. Revista HSM Management. n.11, ano 2 – nov/dez 1998. p.30.

PERKOWITZ, M; ETZIONI, O. Adaptative sites: Automatically learning from user access patterns. Proceedings of the 6th Int. Word Wide Web Conf., Santa Clara, California, April, 1997.

PLANO DIRETOR DA EMBRAPA: REALINHAMENTO ESTRATÉGICO 1999-2003. Brasília: Embrapa-SPI, 1998.36p.

PLANO DIRETOR DO CENTRO DE PESQUISA DE PECUÁRIA DOS CAMPOS SULBRASILEIROS. Brasília: EMBRAPA-SPI, 1993, 42p.

POLÍTICA DE COMUNICAÇÃO. Embrapa, 1996. 57p. (documento para uso interno).

POLÍTICA DE NEGÓCIOS TECNOLÓGICOS. BRASÍLIA: Embrapa-SPI, 1998. 44p. (documento para uso interno).

POPINIGIS, F.; BRANDINI A.; LIMA, S.M.V. e MENDONÇA, S.J.B de Gestão pela Qualidade Total. In: GOEDERT, W.J.; PAEZ, M.L.D'A. e CASTRO, A.M.G. de Gestão em ciência e tecnologia: pesquisa agropecuária. Brasília, EMBRAPA-SPI, 1994.

QUELCH, J. e KLEIN, L. The internet and international marketing. Boston: Sloan Magagement Review, Spring 1996.

QUINLAN, J.R. C4.5: Programs for machine learning. São Mateo: Morgan Kauffman Publishers. 1993.

RESNIK, A. e STERN, B. An analysis of information content in television adversing. Journal of Marketing, January 1977, pp. 50-53.

SAKAMOTO, Y. Tracking web user behavior using event hooks. Proceedings of the WebNet 98 – World Conference of WWW, Internet and Intranet. Orlando-Flórida, Nov., 1-12, 1998.

SALAM, A.F.; RAO, H.R. e PEGELS, C.C. An exploratory analysis of information content of corporate web pages as adversing media.17th International Conference on Information Systems. Dec. 16-18, 1996. Cleveland-Ohio.

SELLTIZ, C. et al. Métodos de pesquisa nas relações sociais. São Paulo: Helder, 1967.

SILVA, J. A. R. e O Marketing na Internet.br – Uma avaliação da presença empresarial na *World Wide Web*. Anais ANPAD 1997. (Edição em CD)

SOARES, R. e HOPPEN, N. Aspectos do uso da *Internet* nos negócios pelas grandes empresas no Brasil: um estudo exploratório baseado em *Sites Web*. Anais ANPAD 1998. (edição em CD)

SPILIOPOULOU, M. e FAULSTICH L.C. WUM: A tool for web utilization analysis. In EDBT Workshop. WebDB'98, Valencia, Spain, Mar. 1998. Available from Word Wide Web: <http://wum.wiwi.hu-berlin.de/wumDescription.html# Publications>

SPILIOPOULOU, M.; FAULSTICH L.C. e WINKLER, K. A data miner analysing the navigational behavior of web users. In Proc. of the Workshop on Machine Learning in user modeling. ACAI'99, Int. Conf., Creta, Greece, July 1999. Available from Word Wide Web: <http://wum.wiwi.hu-berlin.de/wumDescription.html# Publications>

SPINK, A.; BATERMAN, J. e JANSEN, B.J. User´s searching behavior on the Excite web search engine. Proceedings of the WebNet 98 – World Conference of WWW, Internet and Intranet. Orlando-Flórida, Nov., 1-12, 1998.

TAPSCOTT, D. O que esperar do mundo digital. Revista HSM Management n.12, ano 2, jan/fev 1999. p.6.

TAPSCOTT, D. e CASTON, A. Mudança de paradigma: a nova promessa da tecnologia da informação. São Paulo: Makron-McGraw-Hill, 1995.433p.

ZAÏANE, O.R. From Resource Discovery to Knowledge Discovery on the Internet, Technical Report TR 1998-13, Simon Fraser University, August, 1998a. Available from Word Wide Web <http:// www.cs.ualberta.ca/~zaiane/htmldocs/ publication.html>, consulta em 25/02/2000.

ZAÏANE, O.R.;XIN, M. e HAN, J. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs. Proceedings of the Advances in Digital Libraries Conference (ADL'98), Melbourne, Australia, p144-158, April 1998b. Available from Word Wide Web <http:// www.cs.ualberta.ca/~ zaiane/htmldocs/publication.html>, consulta em 25/02/2000.

ZAWISLACK, P.A. Uma proposta de estrutura analítica para sistemas técnicos-científicos: o caso do Brasil. Economia e Empresa, São Paulo, v.3, n.2, p4-29, abr./jun. 1996.

W3C - World Wide Web Consortion. Logging Control In W3C httpd. Available from Word Wide Web <http://www.w3.org/Daemon /User/Config/ logging.html>, consulta em 07/10/1999.

YIN, R.K. Applications of case study reearch. London: Sage Publications, 1993. 129p.

YIN, R.K. Case Study Research: design and methods. 2nd ed. London: Sage Publications, 1994. 171p.