DEMANDA DE INFORMAÇÕES EM WEBSITES: ESTUDO EM UMA EMPRESA DE P&D PARA O AGRONEGÓCIO BRASILEIRO

Ricardo Bernardes [1]
Av. André Tosello, 209 - Barão Geraldo (Campus Unicamp)
CEP: 13083-886 Campinas/SP Brasil
Telefone: (19) 37895802
E-mail: ricardo@cnptia.embrapa.br

Henrique Mello Rodrigues de Freitas [2]
Rua Washington Luis, 855 - Sala 309
CEP 90010-460 Porto Alegre/RS Brasil
E-mail: hf@ea.ufrgs.br

Edimara Mezzomo Luciano [2]
Rua Washington Luis, 855
CEP 90010-460 Porto Alegre/RS Brasil
E-mail: emluciano@ea.ufrgs.br

[1] Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA
Centro Nacional de Pesquisa Tecnológica em Informática para a Agricultura
CEP: 13083-886 Campinas/SP Brasil

[2] Universidade Federal do Rio Grande do Sul - UFRGS
Escola de Administração
CEP 90010-460 Porto Alegre/RS Brasil

Resumo:

A informação é fundamental em qualquer processo de negócio. Neste estudo, busca-se definir e qualificar a demanda por informações em um *website* mantido por uma organização de P&D para o agronegócio. Os dados utilizados consistiram na seqüência de *clicks* realizados por visitantes entre as páginas do *site*, e nas palavras-chave inseridas no seu mecanismo de busca. A discussão dos resultados foi orientada à descoberta de conhecimento que pudesse elevar o nível de personalização e de customização do *site* estudado. Considerando o contexto atual, onde cada vez mais as organizações tenderão a buscar e interagir com clientes e parceiros através da *Web*, este estudo revelou aspectos úteis relacionados às atividades de transferência de tecnologia e de marketing para a organização objeto do estudo.

Palavras-chave: Informação, P&D, agronegócio, website e marketing.

DEMANDA DE INFORMAÇÕES EM WEBSITES: ESTUDO EM UMA EMPRESA DE P&D PARA O AGRONEGÓCIO BRASILEIRO

1 Audiência em *websites:* em busca do *feedback* e da personalização

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 à lenta capacidade dos governos em regulá-la. Isto vem contribuindo também com 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, ao mesmo tempo que tornam a competitividade mais acirrada.

As empresas aumentam cada vez mais sua presença na *Web*: o comércio eletrônico 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 forma, as empresas ficam 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 7 dias por semana.

Considerando esta tendência, tornou-se comum a prática das empresas divulgarem também seus endereços na *Web*. Observa-se que esta tendência abrange todo espectro das organizações, sejam elas privadas, públicas, filantrópicas, comerciais, industriais, de serviços, de ensino, de pesquisa, etc. Ultimamente, a tendência de 'possuir endereço na rede' tem alcançado, também as pessoas físicas, configurada através dos '*sites* pessoais'. Assim, os *websites* têm experimentado significativas melhorias no sentido de tornarem-se adaptativos, agradáveis e chamativos. Tornou-se freqüente 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 a navegação pelos *sites*, tornam-se cada dia mais sofisticados, isto inclui *banners* dirigidos e utilização de *cookies*, entre outros.

O desenvolvimento de ferramentas e técnicas visando minerar padrões de acessos em *sites*, de maneira a personalizar das relações entre o usuário e o *site* acessado tem sido um dos focos de pesquisas acadêmicas recentes (multimídia e inteligência artificial): visam, entre outros, proporcionar ao usuário a recuperação de informações relevantes, investigar qual o seu

comportamento ao atravessar um *Website*, quais são suas estratégias de busca de informação e suas preferências frente aos tópicos apresentados nas páginas.

Dividida entre o direito à privacidade dos usuários e a necessidade de ter *Websites* competitivos, tem-se como principal interessada nestas pesquisas a área de *marketing* das organizações. 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).

Neste contexto, este artigo objetiva relatar um estudo de caso exploratório efetuado em uma empresa de P&D para o agronegócio, o qual visou mapear a demanda por informações requisitadas pelos visitantes em seu *Website*, considerando a freqüência e origem dos acessos, trilhas mais freqüentes entre páginas de conteúdo e as necessidades explicitadas textualmente pelo visitante no mecanismo de busca do *site*. O pressuposto para a realização do estudo foi que os dados poderiam revelar conhecimentos úteis para a adaptação e customização do *site*, visando a personalização das relações com os visitantes (BAMSHAD, 2000), favorecendo diretamente as atividades de *marketing* da organização estudada (BÜCHNER *et al.*, 1999).

Na seqüência, aborda-se a questão dos usuários e seus comportamentos na consulta a um *Website* (seção 2), bem como o método definido para realização desta pesquisa (seção 3), logo após sendo discutidos alguns resultados (seção 4) e apresentadas as conclusões (seção 5).

2 Usuários, suas trilhas e 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-sistema de informação. Ultimamente, com o crescimento do comércio eletrônico, o foco dos estudos está sendo deslocado para a interação usuário-*web*, tendo a área de *marketing* como uma das

principais interessadas.

Num trabalho anterior à 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 desse estudo permitiram a elaboração de uma tipologia de usuários finais, com base em 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*, nos quais 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, semelhante a 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): este elaborou uma biblioteca 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 do *site. 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 à 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 e sistema (*logs*) para posterior análise e utilização.

Em uma pesquisa realizada por JANSEN *et al.* (1998), usando os *logs* do mecanismo de busca *Excite*, foi analisado um *subset* de 51453 consultas efetuadas por 18113 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 31% utilizavam dois termos e outros 31% apenas um termo. Em relação à 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 em 0,23% das consultas 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%, 35%, 66% e 32% 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% 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 *site* adaptativo: 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 *website* 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 do 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 *websites* das organizações. Todas estas técnicas e ferramentas de software estão enquadradas dentro de uma "nova disciplina" denominada *Webmining*. ZAÏANE (1998a) define o termo *Webmining* como sendo a extração de padrões interessantes, potencialmente úteis e de informação implícita de artefatos ou atividades relacionadas com a *World Wide Web*.

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 de *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 artigo, 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 *websites* para indivíduos, sendo voltada para a personalização de *sites* (*Web Personalization*). O assunto vem merecendo a atenção de pesquisadores nas universidades. Dentre as mais expoentes está a Universidade de Minnesota, Universidade Simon Fraser, Universidade de Alberta e Universidade de Ulster.

COOLEY *et al.* (1997a) apresentam o WEBMINER, um software que através da utilização de bancos de dados relacionais e uma linguagem semelhante à 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

website.

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 (estabelecer perfis de usuário, não sendo peculiar a *Web*), e outro relativo ao caminho ou maneira de como a informação é acessada (seu foco é 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.

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.

Evidencia-se 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 *websites*, 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 proporcionar mais foco ao conteúdo, agrupar melhor os conteúdos e recursos, levantar subsídios para profissionais que trabalham com desenho e edição de *sites*, planejar o *site* (estrutura física de pastas, aplicações CGI necessárias, bases de dados, balanço de carga, etc), personalizar *sites*, etc. 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 *websites* sugere, também, formas diversificadas de abordagens para a realização de investigações, favorecendo e valorizando, de forma especial, os estudos de caso de natureza exploratória.

3 Método de pesquisa

Através da literatura apresentada, pode-se inferir que o tema é emergente. Buscou-se apoio em métodos, técnicas e ferramentas da grande área de sistemas de informação para a sua abordagem. A seguir, serão descritos os pressupostos metodológicos e as técnicas empregadas para a condução do estudo.

3.1 Delineamento de pesquisa

O nível de pesquisa considerado mais adequado, conforme GIL (1999, p.45), foi o exploratório; 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 o aprimoramento de idéias ou a descoberta de *insights*. Trata-se, mais especificamente de um estudo de caso: a investigação profunda e exaustiva de um ou de poucos objetos, de maneira a permitir o seu conhecimento amplo e detalhado. Muito embora 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); o estudo de caso apresenta algumas vantagens, tais como o estímulo a novas descobertas, a ênfase na totalidade e a simplicidade dos procedimentos (GIL, 1999).

A unidade de análise foi o *website* de uma unidade descentralizada, de uma empresa de P&D que atua em nível nacional, mas pode-se dizer que foi de fato cada acesso ao *website* considerado. 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): usar múltiplas fontes de evidência, criar uma base de dados do caso e manter uma cadeia de evidência. Pelo fato de ser um estudo simples, procurou-se a convergência das fontes de evidência utilizadas. O relatório seguirá a estrutura linear-analítica, descrita em YIN (1994, p.138).

3.2 Análise dos dados

Foram considerados os seguintes dados e critérios de amostragem:

a) registro de páginas transferidas com sucesso para visitantes não identificados, oriundos do domínio '.br' (domínio reverso resolvido), que acessaram mais de uma página - diferente - durante a visita, no período de 2 de fevereiro de 1999 a 30 de abril de 2000 (453 dias), o que totalizou 26961 acessos distribuídos em 3488 sessões;

b) registro das palavras-chaves inseridas no mecanismo de busca do *site* entre 18h25 de 08/11/1999 e 11h09 de 15/06/2000. A amostra totalizou 2905 termos para busca coletados

em 1473 sessões. Estes foram classificados segundo o *Thesagro* (um manual de catalogação de documentos elaborado pelo Ministério da Agricultura e do Abastecimento e utilizado por documentalistas da área agrícola). Neste conjunto de dados não foi aplicado nenhum critério de seleção, considerando-se todas as consultas realizadas no web*site* no período amostrado.

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 linguagem *SQL-like*, o que permitiu a elaboração de consultas de forma *ad-hoc*. As análises foram conduzidas, considerando-se as estatísticas gerais de acesso ao *site*, as preferências e padrões primários de navegação dos visitantes no *site*, e as preferências explícitas dos visitantes, simbolizadas pelos termos inseridos no mecanismo de busca do *site*. Ao final, elaborou-se um perfil do uso do web*site* e de suas particularidades.

4 Resultados e discussão

São apresentadas a seguir uma análise e discussão das estatísticas de acesso ao *website*, o comportamento ou caminho dos usuários ao navegar pelo *website*, bem como os principais termos usados nas consultas.

4.1. 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.

Em média, o visitante requisitou 5,7 páginas, tendo ficado conectado aproximadamente 8:36 minutos no *site*. Observou-se que 71,9% dos acessos foram provenientes do domínio ".com.br", sendo 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. Estas representaram 2,4% das sessões. As sessões 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).

Apenas 2,8% do total de domínios de terceiro nível registrados, pertencentes ao domínio ".com.br", foram responsáveis por 38,9% das sessões realizadas no *site.* Notou-se 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.

Considerando apenas as instituições de ensino e pesquisa federais ou localizadas fora do Estado do RS, notou-se que apenas 2,6% do total de domínios de terceiro nível diferentes registrados, foram responsáveis por 18,6% do total de sessões realizadas. A UFSM foi a instituição que mais acessou o *site,* com 65 sessões, seguida pela UFRGS, com 63 sessões. Estas instituições apresentaram, também, maior número de *pageviews* por sessão (6,6).

Notou-se, entretanto, que 48% dos acessos foram originados por outras unidades da Embrapa. Uma análise posterior mostrou que as buscas realizadas por outras unidades estava focada na página da equipe técnica. Com relação ao tempo médio de permanência do visitante no *site,* apenas 26,9% das sessões tiveram duração superior a 10 minutos e 28,5% ficaram menos de 2 minutos no *site.* A grande maioria das sessões (73%) transferiu de 1 a 6 páginas e apenas 5% transferiu mais que 15 páginas.

4.2. Os caminhos dos usuários ao atravessar o *site*

Utilizando-se as 2973 sessões que iniciaram a página principal, procurou-se saber quais eram os primeiros *clicks* dos visitantes ao entrar no *site.* As Figuras 1, 2 e 3 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 a notação **A*B*** significa que o visitante iniciou sua sessão na página **A**, acessou - ou não - outras páginas ("*****") antes de chegar à página "**B**", continuando sua sessão através do *site,* ou não ("*****").

A Figura 1 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.

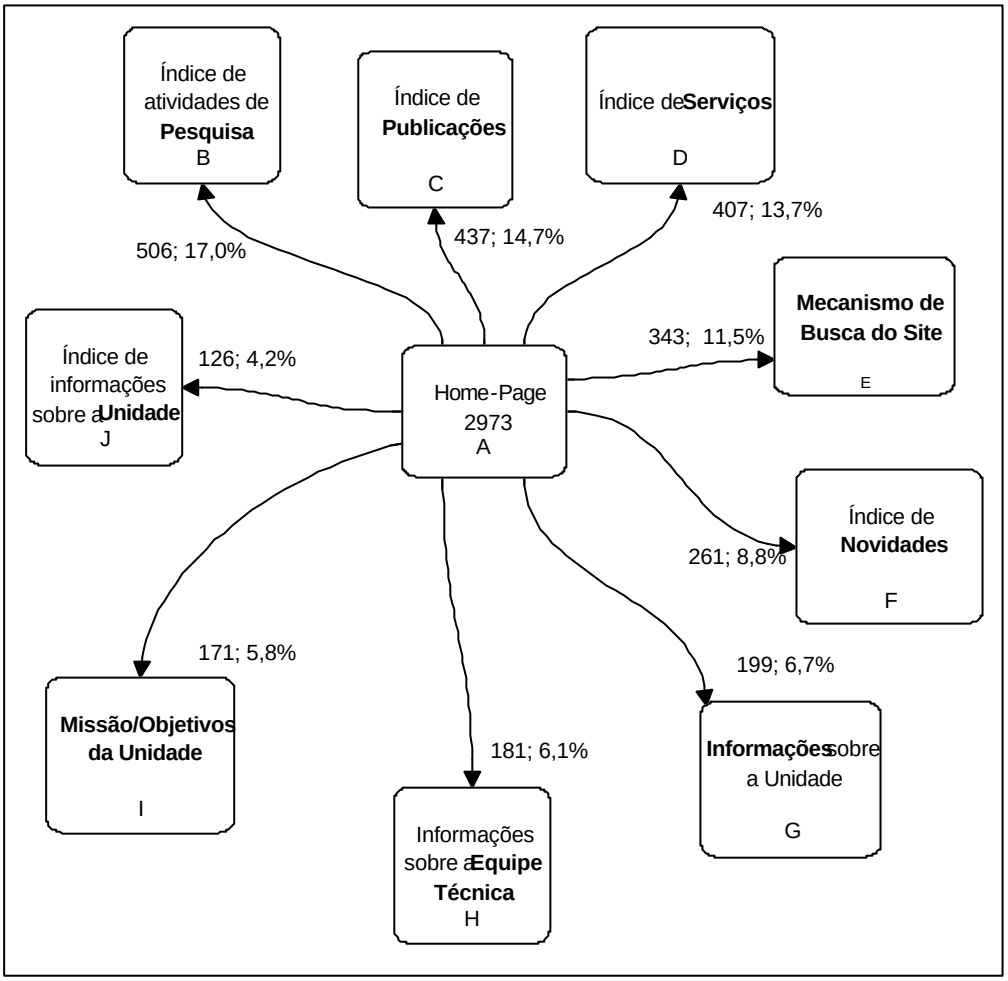

FIGURA 1 - Número de sessões por link acessado na primeira escolha do visitante do *site* (A-B*, A-C*, A-D*, A-E*, A-F*, A-G*, A-H*, A-I*, A-J*)

Visando encontrar regras de preferência considerando o domínio de segundo nível (comercial, governo, instituições de ensino, etc.), os dados da Figura 1, 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 1 a seguir.

TABELA 1 - 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

Os dados da Tabela 1 mostram que havia 72,4% de confiança de que uma sessão, cuja 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*).

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

Nas Figuras 2 e 3 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, reveste-se 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 2, em particular, mostra a primeira decisão dos visitantes cujo primeiro *click* é o *link* "Índice de Serviços". 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).

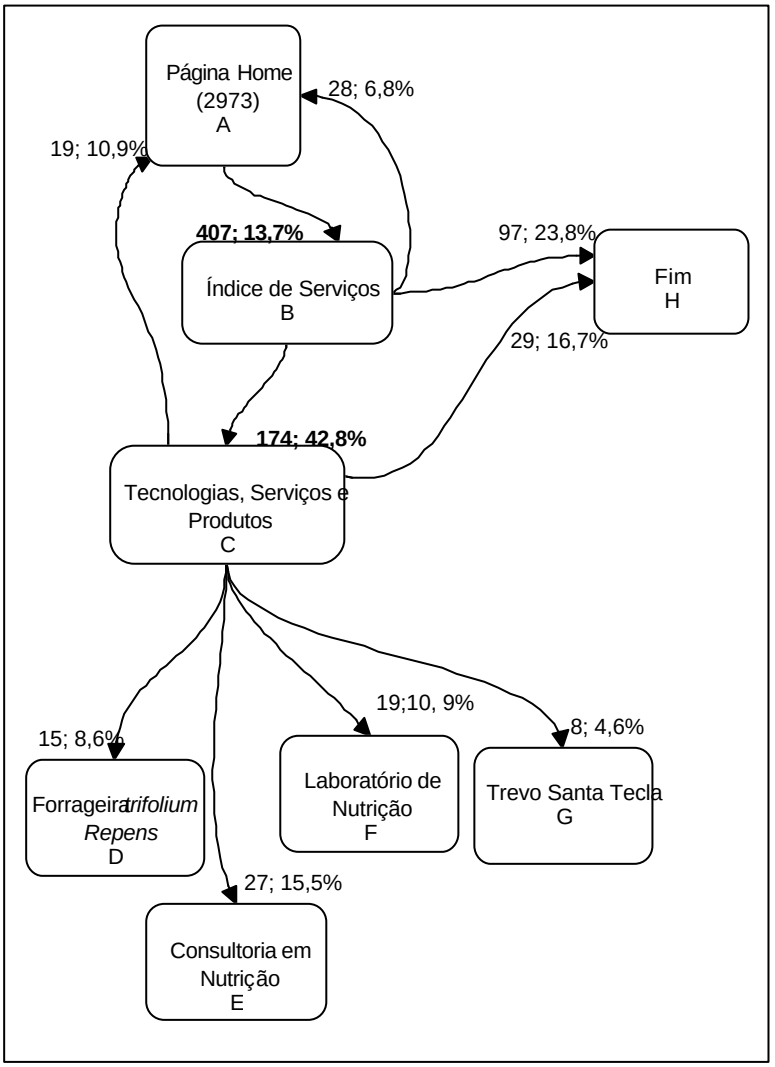

FIGURA 2- Comportamento e preferências dos visitantes que acessaram o *link* "Serviços" na primeira escolha (A-B*, A-B-A*, A-B-C-A*, A-B-C-D*, A-B-C-H, A-B-C-E*, A-B-C-F*, A-B-C-G*, A-B-C-H)

A Figura 3 apresenta o mesmo diagrama da Figura 2, 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 2.

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.

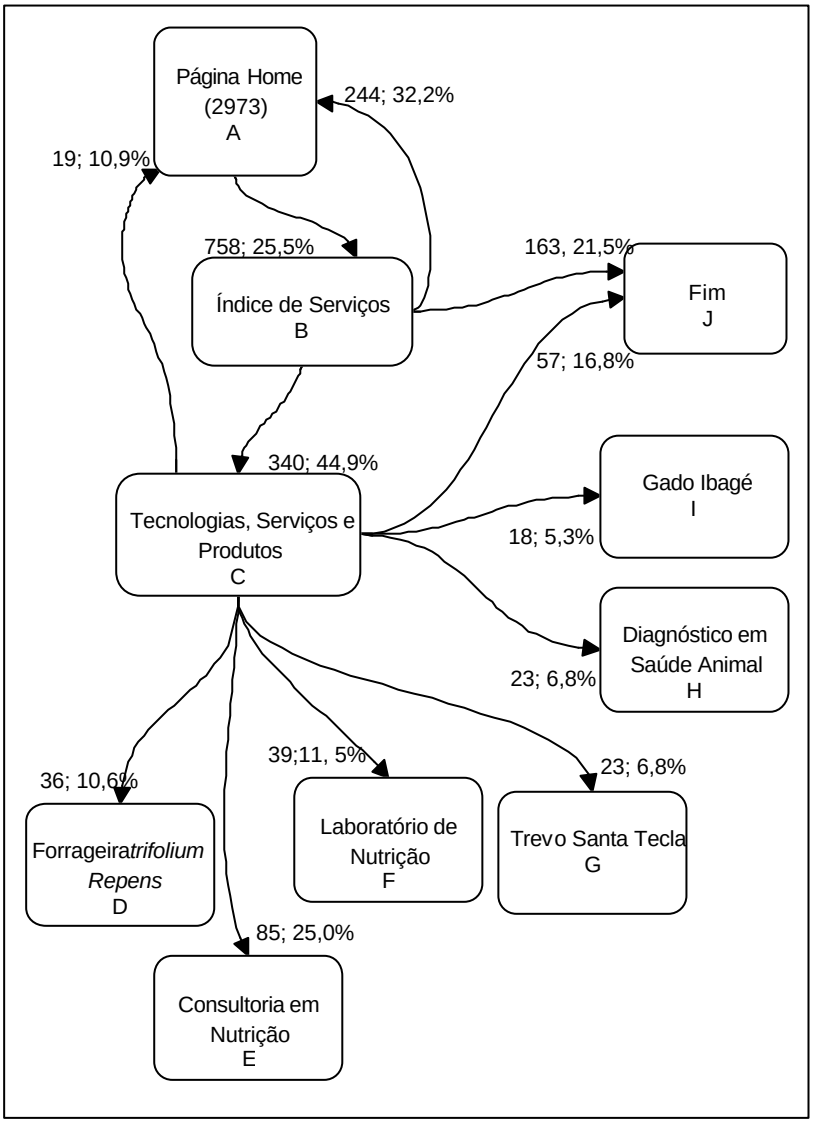

FIGURA 3 - Comportamento e preferências dos visitantes que acessaram o *link* "Serviços" durante a sessão (A*B*, A*B*A*, A*B*C*A*, A*B*C*D*, A*B*C*H, A*B*C*E*, A*B*C*F*, A*B*C*G*, A*B*C*H*, A*B*C*I*, A*B*C*J, A*B*J*)

4.3. Os termos de consulta utilizados pelos usuários do *site*

Em média, o tempo entre consultas na mesma sessão ficou 169,6 segundos. Já o tempo total em que o visitante permaneceu no *site* ficou em 327,4 segundos. 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 as a seguir indicadas.

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, pscicultura, bubalinos, pêssego, etc.

Fora do contexto da unidade e da Embrapa, mas dentro do contexto do agronegócio - termos não cobertos pelas missões da unidade ou da Embrapa ou para os quais a unidade ou Embrapa não se apresentam como referência. Ex.: escargot, 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.

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.

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

Os resultados são apresentados na Tabela 2. 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 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 2 - 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

Com relação ao número de consultas por sessão, notou-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 observou-se que a cada incremento de 1 no número de consultas por sessão, a freqüência da classe cai em média 50%.

Os termos dentro do contexto da unidade, da Embrapa e do agronegócio, totalizaram 2426. Estes foram sumarizados em 1248 termos diferentes. Os termos utilizados 10 vezes ou mais, na exata forma como foram digitados e considerando aqueles contextos, estão apresentados na Tabela 3.

TABELA 3 - 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)

Nota: 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.

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 3 - 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.

A fim de agregar um pouco mais os termos utilizados pelos visitantes, procurou-se classificá-los segundo o *Thesagro*. Dos 2434 termos dentro do contexto da unidade, da

Embrapa e do agronegócio, 72 não foram classificados por aquele sistema, sobrando, então, 2362 termos que foram reduzidos para 435 termos diferentes após a sumarização. Aqueles que apresentaram freqüência igual a 20 ou mais estão apresentados na Tabela 4.

TABELA 4 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)

Notou-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”). Interessante observar o aparecimento do termo “Instalação para Animal”. Isto se deve provavelmente ao fato de que os termos referentes a esta classe variam de região para região, fato que pode apontar para a necessidade de agentes de busca dotados de certa inteligência.

Procurando-se averiguar o interesse dos visitantes focando apenas as espécies animais implícitas ou explícitas na sessão. 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 3 e 4, 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, por exemplo).

Nota-se também que 82,7% das consultas buscavam as espécies animais cobertas pela missão da Unidade (bovinocultura e ovinocultura), devendo-se considerar, que 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 6 é o fato da relação explícita da consulta e implícita na sessão ser muito maior naquelas sessões que buscavam bovinocultura.

TABELA 6 - 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
Suinocultura	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 interessadas em pecuária como um todo.

Através da discussão apresentada, procurou-se enfocar o objeto de estudo (*website*) 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.

5 Considerações finais

Eis algumas considerações a respeito dos resultados obtidos e também sobre alguns aspectos metodológicos deste estudo. Extrapolando o caráter técnico da investigação, resumidamente serão feitas algumas considerações sobre os seus potenciais benefícios, considerando uma perspectiva sócio-econômica.

5.1 Implicações 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 inerentes (no caso) 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 igualmente 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. Apesar de certa igualdade na freqüência de consulta sobre estas duas atividades (412 e 377, respectivamente), notou-se que o *site* é entendido como sendo de bovinocultura de corte. Isto porque, em 61% das consultas em que o visitante buscava informações sobre esta atividade, ela não estava explicitada na *query* (ex.: "reprodução", "nutrição"). Já daqueles que buscavam por ovinocultura, 94,6% explicitaram a atividade no termo de consulta (ex.: "manejo **de ovinos**", "reprodução **de ovinos**").

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 atendê-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 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 possível, 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 aos *softwares* de mineração C4.5 e WUM. Uma versão beta deste último, em particular, foi testado em uma estação Sun Ultra2, com 256MB de memória RAM. Três fatores, entretanto, prejudicaram a tentativa de utilizá-lo: a linguagem utilizada (Java) tem baixa performance, a quantidade excessiva de páginas no site (218), bem como sua estrutura excessivamente profunda produziram resultados confusos, ou, em muitos casos, nem produziram, travando o software. Estas limitações determinaram o abandono do software, e mostraram que o ideal é que o planejamento da estrutura do *site* seja orientado à mineração. Ou seja, a aplicação destas técnicas em *sites* pré-existentes pode se tornar muito difícil.

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 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 *click* - 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, apresentam-se 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*), como os realizados por FREITAS (1993) e 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.

Referências bibliográficas

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.

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.

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.

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

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

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.

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.

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

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.

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.

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.

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.