

## Mineração de informações em portais colaborativos: possibilidade da busca de padrões por meio de algoritmos em portais *wiki*

### *Mining information in collaborative portals: the possibility of using pattern-matching algorithms in wiki portals*

Dani Juliano Czelusniak [dani.juliano@gmail.com](mailto:dani.juliano@gmail.com)<sup>1</sup>

Aline França de Abreu [aline@deps.ufsc.br](mailto:aline@deps.ufsc.br), DSc., UFSC<sup>2</sup>

João Artur de Souza [jartur@gmail.com](mailto:jartur@gmail.com), DSc., UFSC<sup>3</sup>

Gertrudes Aparecida Dandolini [ggtude@gmail.com](mailto:ggtude@gmail.com)<sup>4</sup>

Dario Eduardo Amaral Dergint [dergint@utfpr.edu.br](mailto:dergint@utfpr.edu.br)<sup>5</sup>

<sup>1</sup> Aluno regular do curso de Doutorado em Engenharia de Produção na Universidade Federal de Santa Catarina (UFSC), na linha de pesquisa Inteligência Organizacional sob a orientação da profa. PhD. Aline França de Abreu, onde desenvolve pesquisa voltada ao uso de sistemas agentes híbridos para apoio à processos de tomada de decisão. Possui graduação em Tecnologia em Processamento de Dados pela Faculdade de Ciência e Tecnologia do Paraná (1997), especialização em Bancos de Dados pelas Faculdades SPEI (1998), especialização em Métodos para a Engenharia de Software pela Universidade Tecnológica Federal do Paraná (2002), especialização em Gestão Industrial Conhecimento na linha de pesquisa Conhecimento e Inovação pela Universidade Tecnológica Federal do Paraná (2004), mestrado em Engenharia de Produção pela Universidade Tecnológica Federal do Paraná (2007), Certificações Profissionais Novell CLP - Certified Linux Professional (2008), CLA - Certified Linux Administrator (2008) e LTS - Linux Technical Specialist (2008).

<sup>2</sup> Concluiu o doutorado em Information Technology – Faculty of Engineering – University of Waterloo em 1995. Atualmente é professora de terceiro grau da Universidade Federal de Santa Catarina. Publicou 30 artigos em periódicos especializados e 112 trabalhos em anais de eventos. Possui 2 capítulos de livros publicados. Possui 2 softwares e outro 1 item de produção técnica. Participou de 38 eventos no Brasil. Orientou 83 dissertações de mestrado e co-orientou 2, orientou 12 teses de doutorado, além de ter orientado 6 trabalhos de iniciação científica na área de engenharia de produção. Recebeu 5 prêmios e/ou homenagens. Atua na área de Ciência da Computação, com ênfase em sistemas de informação. Em suas atividades profissionais interagiu com 107 colaboradores em co-autorias de trabalhos científicos. Em seu currículo Lattes os termos mais frequentes na contextualização da produção científica, tecnológica e artístico-cultural são: tecnologia da informação, informação, internet, inovação, sistemas de informação, gestão do conhecimento, tecnologia da informação, competitividade.

<sup>3</sup> Graduação em Matemática (Licenciatura) pela Universidade Federal de Santa Catarina (1989), mestrado em Matemática e Computação Científica pela Universidade Federal de Santa Catarina (1993), doutorado em Engenharia de Produção pela Universidade Federal de Santa Catarina (1999) e pós-doutorado pela Universidade Federal de Santa Catarina (2000). Trabalhou na Universidade Federal de Pelotas de 1993 à 2007 como professor na área de Matemática, atuando também em Educação a Distância. Na Universidade Federal de Pelotas foi coordenador do Curso de Matemática a Distância onde obteve grande experiência com ambiente virtual de aprendizagem, preparação de material didático, e objetos de aprendizagem. Atualmente é Professor Associado da Universidade Federal de Santa Catarina do Departamento de Engenharia do Conhecimento. Na graduação tem trabalhado com disciplinas das áreas de Estatística, Inteligência Artificial e Lógica Matemática. Na pós-graduação tem colaborado com o Programa de Pós Graduação em Engenharia e Gestão do Conhecimento da UFSC na área de métodos de pesquisa quantitativo, Inteligência para Inovação e Gestão das Tecnologias da Informação. Área de pesquisa: Educação a distância, TICs, Inteligência artificial (Redes neurais, Conjuntos difusos, Reconhecimento de padrões), Tecnologias Assistivas no ambiente Moodle, e Inteligência para inovação.

<sup>4</sup> Graduação em Matemática (Licenciatura) pela Universidade Federal de Santa Catarina (1992), mestrado em Engenharia de Produção pela Universidade Federal de Santa Catarina (1997), e doutorado em Engenharia de Produção pela Universidade Federal de Santa Catarina (2000). Foi professora da Universidade Federal de Pelotas entre 2003 e 2007, onde foi coordenadora do curso de Matemática e do curso de Matemática na modalidade a distância. Atualmente é Professora Associada da Universidade Federal de Santa Catarina do Departamento de Engenharia do Conhecimento e pesquisadora da Universidade Aberta do Brasil (UAB). Na graduação tem trabalhado com a disciplina de estatística e inteligência artificial. Na pós-graduação tem colaborado com o Programa de Pós Graduação em Engenharia e Gestão do Conhecimento da UFSC na área de métodos de pesquisa e na área de Mídia. Área de pesquisa: Educação a distância, Inteligência artificial (Redes neurais, Conjuntos difusos, Reconhecimento de padrões), Tecnologias Assistivas, e Inteligência para inovação.

<sup>5</sup> Possui graduação em Engenharia Industrial Elétrica Ênfase Eletrônica e pela Universidade Tecnológica Federal do Paraná (1985), mestrado em *Sciences de L'homme Et Technologie - Université de Technologie de Compiegne* (1996), mestrado em Engenharia Elétrica e Informática Industrial pela Universidade Tecnológica Federal do Paraná (1990) e doutorado em *Sciences de l'Homme e Technologie - Université de Technologie de Compiegne* (1999). Atualmente é professor associado da Universidade Tecnológica Federal do Paraná. Tem experiência na área de Economia, com ênfase em Mudança Tecnológica, atuando principalmente nos seguintes temas: gestão de projetos tecnológicos, inovação tecnológica, sistemas informatizados, empreendedorismo, propriedade industrial e desenvolvimento regional e local.

## RESUMO

Com a popularização da rede Internet e o rápido crescimento tecnológico das ferramentas para publicação de dados em formato hipertexto, surgiram as ferramentas de colaboração para a plataforma web, representadas principalmente pelos *wikis* e *blogs*. Nesse contexto, o objetivo deste trabalho é apresentar uma possibilidade de tratamento para as informações armazenadas nesses portais, que na maioria dos casos não são classificadas e são armazenadas de forma desordenada. As metodologias de pesquisa adotadas são a revisão bibliográfica, na apresentação dos conceitos tangentes ao objetivo deste trabalho e estado da arte na apresentação da forma de classificação e extração de dados. Os resultados demonstram que é possível efetuar mineração de informações para apoio ao gestor em processos de tomada de decisão em informações extraídas desses tipos específicos de portais, após serem tratadas por ferramentas de software específico.

Palavras-chave: Mineração de dados, Portais wiki, Tecnologia.

## ABSTRACT

*With the popularization of the Internet and the rapid growth of technology tools for publishing data in hypertext format, arise the collaboration tools for the web platform, mainly represented by wikis and blogs. In this context, the objective of this paper is to present a possible treatment for the information stored in these portals, which in most cases aren't classified and stored in a disorderly way. The research methodologies adopted are the literature review, for present tangent concepts to the objective of this paper and state of the art in the form of presentation, classification and data extraction. The results shows that it is possible to carry out mining information to support manager in processes of decision-making information from this particular kind of portal, after they were treated by specific software tools.*

*Key-words: Data mining, Wiki portals, Technology.*

## 1. Introdução

Com o advento da popularização da rede Internet, tanto as pessoas quanto as organizações gradativamente, têm modificado a forma de buscar e disponibilizar as informações de seu interesse. Nesse contexto, surgiram tecnologias em software que permitiram a publicação de dados em um formato que poderia ser lido através de navegadores da rede Internet e, mais tarde, esses dados seriam compilados em ferramentas que possibilitariam maior sinergia entre os atores que atuam nesses multimeios. Surge o conceito de “Portais de Internet”.

Com o avanço tecnológico natural, devido ao aquecimento do mercado tecnológico aliado à demanda existente por meios de colaboração, denotado pelo surgimento de ferramentas de comunicação tais como o correio eletrônico e os programas de mensagens instantâneas, a comunidade desenvolvedora de software buscou estruturar uma ferramenta que permitisse maior interação entre as pessoas. Esse trabalho da comunidade resultou no surgimento das primeiras ferramentas focadas em colaboração na web, mais tarde batizadas de *wiki wiki* pela praticidade de seu uso. O termo original, de origem havaiana, significa “muito rápido”.

*WIKI. Neste sistema, grupos virtuais de colaboradores registram e atualizam conhecimentos e práticas de forma rápida e fácil, podendo modificar, escrever, deixar recados ou opinar onde antes podia-se apenas ler (FABRI et al, 2003, p. 10).*

O padrão do *wiki* faz com que a formatação dos dados manuseados, armazenados e publicados muitas vezes não siga a padronização desejada, o que dificulta o trabalho de

categorização e resgate desta informação para posterior tratamento e utilização em apoio ao gestor em processos de tomada de decisão. Isso decorre do processo de colaboração, em que a ferramenta se propõe a permitir que as pessoas possam interagir no ambiente da rede internet, buscando melhorar o desempenho das atividades executadas através de conceitos de colaboração. Nesse sentido, a colaboração em ambiente web implica muitas vezes na não padronização do dicionário utilizado.

Nesse caso, uma das possibilidades seria o uso de algoritmos para a organização e identificação de padrões desses dados através da aplicação de modelos que permitam a mineração de informações, ou *datamining*, seguindo dicionários de dados e posterior categorização das informações em estruturas lineares indexadas por categorias e relações de proximidade.

*O conceito de datamining está justamente associado à esta necessidade de se buscar dados de forma temporânea e coerente, em fontes externas e não estruturadas de informações. Isto é feito mediante o uso de métodos e softwares que permitam uma adaptação imediata ao caráter fugaz e mutante do contexto informativo do mercado (DORNELAS et al., 1999, p. 09).*

Para verificar essa possibilidade, este artigo busca explicar sobre a viabilidade desse processo, observando a forma por meio da qual as informações são armazenadas e interligadas em um portal do tipo *wiki* e de que forma uma ferramenta que permita esta mineração poderia chegar aos resultados desejados.

## **2. A tecnologia da informação nas organizações**

No contexto de evolução tecnológica em que a sociedade encontra-se imersa, se observa o crescente uso de ferramental tecnológico pelas organizações, que buscam aumentar

seu potencial competitivo frente às demandas de mercado, reduzindo seu tempo de resposta a incidentes e baixando o seu custo de operação. Notoriamente, as organizações modernizam suas estruturas através da adoção de ferramentas que permitam a otimização de seus processos empresarias.

Os sistemas informatizados utilizados hoje pelas organizações surgiram como a evolução dos Sistemas MRP (*Manufacturing Resource Planning*) ou Sistemas de Controle de Manufatura. Nas palavras de Senapeschi Neto e Alliprandini (1998) o modelo MRP surgiu devido à necessidade dos gestores em ter um ferramental de apoio para o processo de gestão e de cálculo das necessidades de materiais. Em seguida, com a melhoria do modelo, surgiu o MRPII, que incorporou as necessidades de capacidade e outros recursos que, junto aos materiais, possibilitaram sustentar a gestão da manufatura como um todo.

Devido ao sucesso da automação dos processos de manufatura, os gestores se questionaram sobre o motivo pelo qual as demais atividades empresariais, que implicam diretamente ou indiretamente no processo de manufatura, até então, não poderia também ser gerenciada com o auxílio de ferramental de software. Dessa forma, surgiu a demanda para a construção dos sistemas de gestão organizacionais integrados. Esses sistemas de gestão são comumente tratados pelo mercado como sistemas ERP (*Enterprise Resource Planning*) em sua definição original, citado pelos autores nacionais como SGE (Sistema de Gestão Empresarial).

Hoje, com o advento da universalização da rede internet, observada no contexto organizacional por sua utilização pelos usuários finais, estes são tratados pelas organizações

sob o escopo de clientes internos (ou clientes em potencial). Novamente as organizações voltam seus olhos para as novas tecnologias baseadas nos padrões da Internet.

### **3. Portais Cooperativos: novo paradigma da informação organizacional**

A rede Internet se estabeleceu no Brasil nos idos de 1990, utilizada inicialmente para fins de pesquisa entre as Universidades Federais, órgãos governamentais e algumas empresas de comunicação estatais. Com o surgimento dos “provedores de acesso” à rede, discretamente as empresas aderiam à Internet com a finalidade de utilizá-la apenas como veículo de propaganda em formato eletrônico.

Com o passar dos anos e o amadurecimento da tecnologia empregada, observou-se o crescimento da rede Internet. Pode-se notar o surgimento de novas formas de interação entre os indivíduos presentes no ambiente virtual da rede, também tratado como *cyberspace* ou ciberespaço. Esse termo foi criado pelo escritor Willian Gibson em sua obra *Neuromancer* e hoje é amplamente utilizado quando há referência ao ambiente virtual da rede internet.

O amadurecimento tecnológico da rede culminou em novas tecnologias que modificaram o comportamento das pessoas, que de um modo geral utilizam a rede para fins pessoais e profissionais. Dentre as tecnologias voltadas a interação, suas vertentes apontam para as tecnologias sedimentadas no novo padrão Web 2.0, que permite o desenvolvimento de tecnologias baseadas em portais com mecanismos que enriquecem o modelo de interação entre as pessoas. Com o padrão Web 2.0, algumas aplicações ganharam rapidamente espaço na rede Internet, como as redes de relacionamento *Orkut* e *Facebook*, mecanismos baseados em portais que permitem o trabalho cooperativo como os *wikis*, *blogs* e *microblogs*, troca de

mensagens instantâneas baseadas em salas de conversação também chamadas de *chat rooms*, dentre outras.

*Portais e Serviços Web: A terceira camada da plataforma é composta pelos instrumentos desenvolvidos para apresentação de informações na Web (websites), para publicação de informações dinamicamente atualizadas com interação com a comunidade usuária e pelos recursos de disseminação de serviços de informação de governo na Web (Web services) (ZAMAI et al, 2004, p. 02).*

Rapidamente o modelo de interação entre as pessoas proporcionado por estas novas tecnologias, aliado à sua curva exponencial de aceitação dentre os mais diversos grupos sociais, foi identificado pelo mercado, culminando em que diversas empresas iniciassem seu uso, buscando estreitar suas relações com o mercado consumidor. Atualmente, conseguimos observar empresas divulgando seus produtos através de portais específicos para o compartilhamento de vídeos como o *You Tube*, consultores prestando esclarecimentos a grupos de consumidores reunidos no portal da rede de relacionamentos *Orkut*, *headhunters* de empresas especializadas em realocação de executivos efetuando buscas em portais como o *facebook*, dentre várias outras possibilidades que o ambiente da rede internet hoje proporciona.

Os portais da rede internet são ferramentas bastante arrojadas tecnicamente. Possuem recursos e tecnologias embutidas que facilitam a interação entre as pessoas, buscando proporcionar experiência de navegação mais agradável e ágil. Sob o ponto de vista da usabilidade, as organizações procuram melhorar suas tecnologias em portais para que o cliente consiga de forma agradável, encontrar a informação desejada no menor tempo possível.

*Dentre as diversas ferramentas enquadradas na tecnologia da informação, ressalta-se o desenvolvimento de portais na Internet, que proporcionam aos seus visitantes, interação fácil e em tempo real, ou seja, permite que se acesse o que estiver acontecendo em qualquer parte do mundo, diminuindo assim as distâncias e o tempo (FREITAS et al., 2001, p. 06).*

Estruturalmente, a base tecnológica necessária para a implantação de um portal Internet é bastante semelhante à utilizada por um sistema de gestão tradicional. As informações são disponibilizadas e manipuladas pelos visitantes e clientes através de sua interface gráfica (que neste caso ocorre através do navegador de Internet ou *browser*). Posteriormente, os dados são armazenados em bases de dados de onde podem ser resgatados diversos tipos de informações. Com o aumento do volume de dados armazenado aliado ao amadurecimento de seu modelo de negócio e de gestão de informações, existe a possibilidade da extração de informações estatísticas, por exemplo, padrões de navegação, perfis de consumidores e de consumo, dentre outras.

#### **4. Minerando padrões através de algoritmos de busca de proximidade**

A mineração (ou prospecção) de dados, prática conhecida por seu termo original do idioma inglês *data mining*, é um tipo de busca de informações em grandes volumes de dados, geralmente bases de dados históricas, em que a informação nem sempre está formalmente catalogada e organizada. Seu objetivo é efetuar de forma analítica a coleta de informações direcionada à tomada de decisões, buscando através de algoritmos por relações nos dados coletados. Nesse contexto, diversos tipos de algoritmos de seleção e análise podem ser utilizados no processo de mineração de dados. Dentre as possibilidades, podem ser citadas as

árvores de decisão, redes neurais, modelos estatísticos diversos como os de regressão e correlação, dentre diversos outros.

Entretanto, devido à característica distribuída da informação presente nos portais da Internet baseados nos conceitos de Web 2.0, aliada a sua forma ímpar de armazenamento de conteúdo não estruturado, vislumbra-se a possibilidade da utilização de ferramentas que permitam efetuar o tratamento de conceitualizações e terminologias compartilhadas entre as pessoas que interagem através de um portal. Essa forma de agrupamento e mineração de informações é conhecida como *folksonomia*, ou taxionomia de proximidade. Esse termo foi cunhado por Thomas Vander Wal, no ano de 2004, após sites como o *Flickr* e *del.icio.us* utilizarem marcações de rótulos (*tags*) definidas por seus usuários para categorizar informações. Uma página estática publicada na rede Internet pelo próprio autor<sup>i</sup> demonstra o processo de surgimento e rotulação do termo.

A *tag* indica a classificação que determinado usuário atribui para o dado que está sendo publicado, independentemente do contexto em que está inserido. Por exemplo, um *wiki* qualquer pode possuir as mais variadas categorias de informações postadas. Cabe ao moderador do portal filtrar (ou não) o conteúdo que é postado pelos usuários através de regras de moderação.

Utilizando essa *tag* como índice de busca para a carga de dados relativos ao conteúdo de páginas, pode-se através de estruturas articuladas a partir de dicionários de dados, atribuir relações de proximidade entre palavras e frases, indicando a possível categoria em que se insere o conteúdo postado. Pode-se utilizar como exemplo as palavras informática, computador, impressora, papel e sistema operacional. Estabelecendo-se relações através de

dicionários de termos que se deseja categorizar e relacionar, conforme demonstra a Figura 1, a palavra “papel” pode estar inserida em dois contextos distintos, um de informática como insumo de impressoras e em outro prisma como artigo de papelaria.

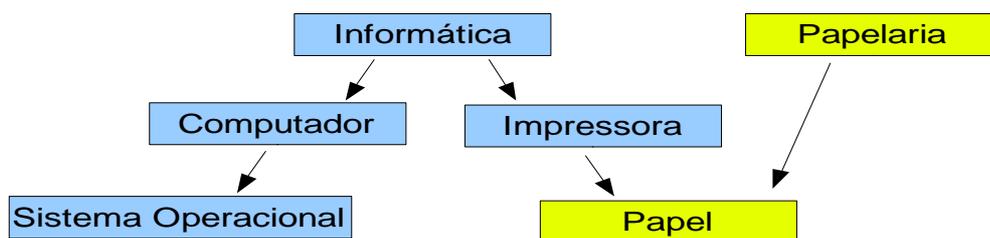


Figura 1: Categorização de palavras

Definida a estrutura para relacionamento das palavras utilizadas como *tags*, é possível efetuar correlações na estrutura, permitindo auxiliar a identificação de modelos para observação de agrupamento de informações para apoio à processos de gestão e de tomada de decisão.

## 5. Possibilidade de mineração de informações em portais baseados em *wiki*

A terminologia *wiki* é utilizada para referenciar um tipo específico de portal focado em trabalho colaborativo, que tem como premissa básica disponibilizar informação manuseada por um grupo, equipe ou comunidade em formato hipertexto. Seu uso hoje, com a popularização da rede Internet é bastante difundido principalmente na estruturação de portais em que ocorre a interação de pessoas na forma de comunidades virtuais com os mais variados objetivos.

Existem hoje diversas ferramentas de software para a implementação de portais baseados em estruturas *wiki* disponíveis na rede Internet. Rapidamente, pode-se citar o

*mediawiki* que é a ferramenta utilizada no site colaborativo *Wikipedia*, *dekiwiki*, *tikiwiki*, dentre muitos outros. Da mesma forma, com o que ocorre em diversas outras ferramentas baseadas em portais (também nos sistemas de informação tradicionais), as informações manuseadas por um site do tipo *wiki* são armazenadas em um sistema gerenciador de banco de dados.

*O Sistema Gerenciador de Banco de Dados (SGBD) é um software onde os dados armazenados por sistemas aplicativos podem ser manipulados através de uma interface de acesso, garantindo confiabilidade e integridade nos processos de armazenamento e recuperação dos dados através de mecanismos chamados de transação. No mercado os sistemas gerenciadores de grande porte mais conhecidos são os produtos Oracle, SQL Server da Microsoft, DB2 da IBM e o software livre PostgreSQL (CZELUSNIAK, 2007, p. 51).*

Existem diversas ferramentas no mercado que se propõe a esse fim, ferramentas livres de ótima qualidade mantidas pela comunidade e soluções proprietárias que são adquiridas sob licenciamento de software específico. Comumente, as ferramentas *wiki* disponíveis na rede Internet são licenciadas sob o modelo de software livre (que são as mais comuns e mais difundidas), são compatíveis com sistemas gerenciadores de bancos de dados que, na mesma forma, são distribuídos como software livre. Como exemplo, podem ser citados o PostgreSQL e o MySQL. Em sua grande maioria, essas ferramentas seguem o padrão de desenvolvimento LAMP.

Nas palavras de Fujisaki *et al.* (2004), o padrão LAMP é um conjunto de tecnologias baseadas em software livre que possibilita a obtenção de sistemas de alto desempenho e baixo custo pelo uso dos softwares abaixo listados:

- **Linux:** Sistema operacional para computadores.
- **Apache:** Servidor de aplicações compatível com os padrões da Internet.
- **MySQL:** Sistema gerenciador de banco de dados simplificado.
- **Php:** Linguagem baseada em scripts bastante comum no ambiente da Internet.

Então, toda a informação manuseada por um internauta quando da interação com um portal baseado em *wiki* é depositada nesse sistema que gerencia os dados manipulados. Para o início do processo, é necessário que o internauta se cadastre; os dados necessários para este cadastro podem variar de acordo com as informações que o administrador do site quiser armazenar. Após esse cadastro, pode-se postar conteúdo no portal, de acordo com a forma de interação que este novo utilizador possuir. Essa forma de interação varia de acordo com papéis atribuídos pelo moderador do portal, variando desde o usuário convencional até papéis de administrador, que possui maior nível de permissões para efetuar modificações estruturais no *wiki*.

A mineração de informações na base de dados desses portais pode ser estruturada com a mesma mecânica através da qual são colhidos dados em sistemas de informação convencionais como sistemas de gestão empresariais. É necessário se definir o que se precisa buscar, então, construir instruções em linguagem SQL<sup>ii</sup> para localizar e quantificar as informações armazenadas desejadas.

Porém, nos sistemas de informações convencionais os dados são armazenados de acordo com uma estrutura específica chamada de “modelo de dados”, que é a forma definida

pelos desenvolvedores da aplicação, que indica onde a informação relativa a cada processo deverá ficar armazenada. Por exemplo, os dados relativos aos clientes estão armazenados na área da base de dados designada às informações dos clientes; os dados das vendas na área das informações sobre vendas etc. Sua padronização ocorre através da construção de modelos chamados de Diagramas de Entidade e Relacionamento (DER) e Diagramas de Fluxo de Dados (DFD).

*Os modelos de requisitos existentes descrevem um ambiente organizacional em termos de entidades e atividades, sem se importarem com situações em que os usuários poderão tomar diferentes decisões. Esses modelos têm como objetivo considerar a descrição de sistemas técnicos e não para fornecer descrições mais ricas sobre as organizações sócio-humanas. As informações capturadas nos modelos existentes, como Diagrama de Fluxo de Dados (DFD) e Diagrama Entidade e Relacionamento (DER), não são suficientes, uma vez que esses modelos descrevem apenas entidades, funções, fluxo de dados e estados do sistema, não expressando as razões envolvidas no processo, ou seja, o porquê fazer uma determinada ação ou tomar uma decisão (PÁDUA e CASARINI, 2001, p. 03).*

Quando se trata de dados armazenados em portais *wikis*, a base de dados denota apenas que aqueles registros são respectivos às páginas de um portal, indicando no máximo uma categoria (se existir) para aquela informação, o texto publicado e onde estão as eventuais imagens que fazem parte do texto. Não existe organização modelada das informações ali contidas, o que dificulta (e algumas vezes impossibilita) o aproveitamento dessas informações em processos empresariais. Para o pleno aproveitamento dessas informações, faz-se necessária a composição de uma estrutura baseada em filtros, que permita saber o que foi armazenado na base de dados do portal. Posteriormente, os indícios capturados pelos filtros precisam receber tratamento qualitativo e quantitativo para classificação. Em um terceiro

momento, precisarão ser definidas as relações de proximidade entre esses itens, indicando o grau de proximidade e relação existente entre os termos capturados.

## **Considerações finais**

Os portais de uso colaborativo baseados em *wikis* e *blogs* ganharam notoriedade na grande rede e rapidamente passaram a ser utilizados pelas corporações, buscando aumentar a sinergia entre seus clientes externos e internos. Logo foi percebida a riqueza do “complexo ecossistema” de informações armazenadas em suas bases de dados, que pode conter todo o tipo de informação possível, desde informações de clientes sobre produtos e serviços até indicativos de necessidades presentes e futuras etc. Tudo depende de para qual finalidade a ferramenta é utilizada. Porém, como o portal é um software que opera de forma distribuída, possibilitando aos atores sua interação através da inserção de informações não estruturadas, o processo de mineração dessas informações pode se tornar uma tarefa complexa.

Tecnicamente, pode-se afirmar que as informações estão armazenadas de forma pseudo-ordenada pelo uso das *tags* na base de dados. Uma das formas através da qual se pode efetuar o mapeamento de suas informações é montando estruturas semânticas em linguagem SQL através do uso de cláusulas como *distinct*, *like* e *count(\*)*, identificando e quantificando ocorrências dos termos armazenados previamente. Posteriormente, esses termos deverão ser hierarquicamente ordenados e categorizados na forma de um dicionário de dados. Com essas informações em mãos, pode-se atribuir pesos às palavras constantes no dicionário de dados, bem como definir suas relações de proximidade entre os termos encontrados. Com esse processo, obtém-se como resultado um “modelo em árvore” em que os termos com maior

grau de ocorrência aparecerão no topo da árvore, associados aos termos em menores graus de ocorrência nos ramos capilares da estrutura.

Como toda a informação armazenada está sempre ligada ao usuário cadastrado no portal que a inseriu ou modificou, pode-se facilmente relacionar o dado com o autor. Então, o filtro pode ainda ser melhorado, se nas informações cadastradas pelo autor constarem suas áreas de interesse, formação ou termos-chave que indicam que tipo de informação aquele autor costuma gerar. Dinamizando ainda mais o algoritmo, podem ser estipuladas estruturas baseadas na competência declarada pelo autor em seus dados cadastrais, com um dicionário de dados elaborado a partir dos termos encontrados nos textos que o autor disponibiliza no próprio wiki ou *currículo* publicado sob um formato, como o da plataforma Lattes, por exemplo. Dessa forma é possível conseguir mensurar níveis de qualidade e confiabilidade à informação publicada, efetuando correlações entre a publicação e os dados presentes em sua plataforma curricular.

De posse dessas informações, é possível elaborar um modelo de qualidade para postagem dentro da ferramenta *wiki*, em que caso o texto redigido pelo autor contenha os termos associados à sua competência ou o que este costuma publicar, o referido texto é publicado sem a necessidade de moderação por parte do administrador do site. Caso o contrário, pode-se fazer com que o material passe pela validação de autores que possuam aquelas competências apontadas pelos filtros declaradas em seus dados pessoais ou quantificadas através de suas publicações.

---

<sup>i</sup> <http://vanderwal.net/folksonomy.html>

ii A linguagem SQL (*Structured Query Language*) ou Linguagem de Consulta Estruturada é um tipo de linguagem padrão para operações com sistemas gerenciadores de bancos de dados.

## Referências bibliográficas

CZELUSNIAK, Dani J. **Proposta de sistema para apoio à gestão de competências baseada em sistemas agentes**. 2007. Dissertação. Universidade Federal Tecnológica do Paraná. Ponta Grossa. PR.

DORNELAS, Jairo S.; BARROS NETO, José P.; OLIVEIRA, Jessé G. **Potencialidades estratégicas associadas ao uso de *datamining***. 1999. Artigo. XIX Encontro Nacional de Engenharia de Produção. Rio de Janeiro. RJ.

FABRI, Adriano; HOLZBACH, Rafael H.; FURUKAVA, Johnny; GAUZA, Alexandre. **Ambientes virtuais de aprendizagem e a criação do conhecimento organizacional**. 2001. Artigo. IV ISKM/DM - International Symposium on Knowledge Management / Document Management. Curitiba, PR.

FREITAS, Maria C. D.; LIMA, Luciana M. S.; CASTRO, João E. E. **Aplicação das novas tecnologias para seleção da informação no setor da construção civil**. 2001. Artigo. XXI Encontro Nacional de Engenharia de Produção. Salvador. BA.

FUJISAKI, Sílvio T.; FRÈRE, Annie F.; SCHLUNDZEN JÚNIOR, Klaus; SCHLUNDZEN, Elisa. **Desenvolvimento de Ferramentas para Páginas web com Recursos de Acessibilidade para Pessoas com Necessidades Especiais**. 2004. Artigo. IX Congresso Brasileiro de Informática em Saúde. Ribeirão Preto. SP.

PÁDUA, Sílvia I. D.; CAZARINI, Edson W. **Modelagem organizacional, facilitador do desenvolvimento de sistemas de informação**. 2001. Artigo. XXI Encontro Nacional de Engenharia de Produção. Salvador. BA.

SENAPESCHI NETO, Alberto; ALLIPRANDINI, Dário H. **Raciocínio Sistêmico e Gerência de Processo: Uma Proposta para Identificá-los nos Sistemas de Gestão Integrada da Produção**. 1998. Artigo. XVIII Encontro Nacional de Engenharia de Produção. Niterói. RJ.

ZAMAI, Kristiany K.; MARCHEZAN, Marcos L.; RISO, Bernardo G.; PACHECO, Roberto C. S. **Diretório de Instituições: Framework para a gestão de uma unidade de informação sobre o sistema nacional de Ciência, Tecnologia e Inovação**. 2004. Artigo. XXIV Encontro Nacional de Engenharia de Produção. Florianópolis. SC.