CSCOS – Organização de Conteúdo Orientada por Senso Comum

Wanderley S. Wang e Lucia Filgueiras
Instituto de Pesquisas Tecnológicas do Estado de São Paulo (IPT)
Av. Prof. Almeida Prado, 532 São Paulo - SP Brasil
+55.11.3767-4068
wanderleywang@yahoo.com.br; lfilguei@gmail.com

ABSTRACT

This paper presents the CSCOS (Common Sense Context Organization Scheme), an organization scheme for web sites built using the Common Sense Knowledge Base of project OMCS-Br (Open Mind Common Sense Brazil). A case study of application of CSCOS shows the similarities and differences between this approach and organization built using card-sorting, a traditional approach.

RESUMO

Este trabalho apresenta o CSCOS (*Common Sense Context Organization Scheme*), um esquema de organização de conteúdo de *sites* com base no contexto detectado com apoio de uma Base de Conhecimento de Senso Comum, especificamente naquela gerada pelo projeto OMCS-Br (*Open Mind Common Sense* no Brasil). Um estudo de caso de aplicação do CSCOS mostra as semelhanças e diferenças entre essa abordagem e a organização por *card-sorting*.

Keywords

Arquitetura da Informação, Senso Comum, Encontrabilidade, Interação Humano-Computador, Card Sorting, Similaridade de contexto.

INTRODUCÃO

Categorizar é agrupar entidades (objetos, idéias, ações, etc.) por semelhança. Essa é uma habilidade natural que a mente humana usa para compreender o mundo ao seu redor. Em geral, um *site* com razoável volume de conteúdo é organizado em seções, as quais podem ter um ou mais níveis de subseções. As características comuns dos itens de conteúdo definem o esquema de organização, influenciando o agrupamento lógico desses itens.

Projetar o sistema de organização para *sites* com grande volume de informação (muitos itens de conteúdo) não é tarefa trivial, principalmente por serem inúmeras as opções de agrupamento. Se a categorização não for eficaz, é possível que o usuário não encontre o que quer.

IHC 2010 – IX Simpósio sobre Fatores Humanos em Sistemas Computacionais. October 5-8, 2010, Belo Horizonte, MG, Brazil.

Para incentivar um usuário a usar e participar de um *site*, uma das melhorias técnicas mais evidentes que podemos fazer é prover uma boa arquitetura de informação, na qual a forma de estruturação de conteúdos colabore para que os usuários tenham uma melhor experiência.

É razoável pensar que seria útil se o projetista tivesse meios de aproveitar o conhecimento social das pessoas para ajudá-lo a organizar o conteúdo de um *site* de uma forma que os usuários achem coerente e familiar. Em um trabalho anterior [28], propusemos relacionar os estudos da disciplina de Arquitetura da Informação aos estudos sobre Bases de Conhecimento de Senso Comum, da área de Inteligência Artificial, para propor um processo de geração do sistema de organização para *sites* que possuem muitos itens de conteúdo textual (ou seja, páginas web).

Esse processo, que nomeamos CSCOS (*Common Sense Context Organization Scheme*), consiste em produzir uma organização que leva em conta o contexto do assunto tratado pelos textos do *site*, contexto esse detectado com apoio de uma Base de Conhecimento de Senso Comum.

Acreditamos que uma organização que leve em conta o contexto pode ser favorável a encontrabilidade no caso de uma pesquisa exploratória. Neste artigo, expomos com detalhe o processo desenvolvido e apresentamos os resultados de sua aplicação a um *site*, comparando a organização produzida com a gerada pelo processo tradicional de *card-sorting*. Para tanto, nas próximas seções deste trabalho apresentamos: (i) o problema que nosso processo pretende tratar e os fundamentos teóricos que embasam esse estudo; (ii) detalhes do funcionamento do processo; (iii) um estudo de caso que utiliza esse processo; e, por fim, (iv) nossas conclusões e considerações finais.

FUNDAMENTOS TEÓRICOS

Organizar uma grande quantidade de itens de conteúdo (por exemplo, textos diversos) é algo que pode ser feito isoladamente pelo editor ou projetista do *site*. Mas um bom projeto deve ser centrado no usuário, conforme apontam diversas áreas de estudo, como Interação Humano-Computador (IHC), *Human-Information Interaction* (HII), *Human-Centered Computing* (HCC) ou UX (*User eXperience Design*). Há uma considerável quantidade de pesquisas sobre como organizar de maneira eficiente o conteúdo de um *site*, observando o que as pessoas

conhecem e como se comportam no mundo real. A seguir sumarizamos uma pequena parte desses trabalhos.

Organização de Conteúdo e a Pesquisa Exploratória

A Arquitetura da Informação (AI) é uma disciplina que orienta o projetista (arquiteto da informação) sobre como estruturar um espaço informacional de modo a facilitar o acesso intuitivo das pessoas aos conteúdos existentes, por exemplo, em projetos de *sites* [26].

Seus componentes principais, os sistemas de organização, rotulação, navegação e de busca, formam os pilares básicos de um *site*, elaborados levando-se em conta o relacionamento e a natureza de interdependência existente entre usuários (público-alvo), conteúdo (documentos, tipos de dados, serviços) e contexto (objetivos do negócio, vocabulário, políticas) dentro de uma complexa e adaptativa ecologia da informação [21].

Um **sistema de organização** estrutura o conteúdo existente para que as pessoas possam encontrar a informação que procuram. Morville [20] sugere que a qualidade de um *site* possa ser avaliada pela análise de diversas facetas, sendo uma delas a encontrabilidade ("*findability*", em tradução livre), que se refere ao grau de facilidade com que determinada informação é localizada e que pode ser priorizada pelo projetista para melhorar a experiência que os usuários terão ao visitar um *site* [19].

A dificuldade para se encontrar algo pode advir de vários fatores: por exemplo, de uma organização confusa por falta de padrões, da ambigüidade natural existente na linguagem, das diferenças de perspectivas entre os criadores da organização e seus usuários, ou até mesmo da falta de familiaridade do usuário no *site*. A incapacidade de encontrar uma informação é um dos fatores que mais desagradam os usuários, e corrigir esse problema economiza tempo e aumenta sua satisfação [21].

Uma das tarefas mais afetadas em função da organização do conteúdo é a pesquisa exploratória, ou seja, a que envolve a coleta de múltiplos pedaços de informação para entendimento de um assunto ou tomada de decisões. Segundo Pirolli [25], a pesquisa exploratória ("*berry picking*") é o tipo de tarefa mais freqüente (71%) dentre as que as pessoas consideram como importantes feitas por elas na internet. A busca por informações específicas responde por 25% da freqüência, comparativamente.

Diversos autores estudaram o comportamento das pessoas enquanto procuram informação. O *Sense Making*, por Brenda Dervin [8] e o ISP (*Information Search Process*) [14], mostram como as emoções e incertezas afetam a qualidade e a natureza da pesquisa e dos seus resultados. Outro aspecto importante é que, segundo alguns autores [4, 25], as pessoas atuam sob "racionalidade limitada" ("*bounded rationality*"), com limites de tempo e de recursos. Numa pesquisa exploratória, as pessoas, em geral, estão diante de grande volume de informações e da necessidade de tomar decisões complexas; portanto, com tempo limitado, tendem a se contentar com uma solução "suficientemente satisfatória". Por isso, se a organização adotada dificulta sua capacidade para encontrar informações relevantes, os usuários abandonam o *site*.

O comportamento das pessoas enquanto procuram informação também é o foco da Teoria do Forragear [25]. Forragear significa "vasculhar, remexer, à procura de algo" e, também, a atividade de "um ser vivo procurar alimento, lançando mão de estratégias especializadas, desenvolvidas no âmbito da espécie" [9]. Trata-se de uma abordagem que, usando diversas analogias, estuda as estratégias que os usuários usam para "caçar" a informação através da análise de "pistas", ou do "cheiro da informação". O conceito de "cheiro da informação" ("*Scent of Information*") é uma alusão ao odor que os animais sentem pelo faro quando estão caçando uma presa. O cheiro da informação é produzido por sinais (como o texto do rótulo do *link*, a grafia da URL ou textos e figuras próximas) associados aos *links*, que os usuários usam de forma subjetiva para prever a possível utilidade do conteúdo distante [7, 27]. Portanto, para orientar os usuários em suas pesquisas, o projetista precisaria lançar mão de palavras-gatilho: palavras ou frases, existentes nas páginas do *site*, que o usuário associa a um conteúdo e influenciam suas escolhas de navegação, com maior ou menor força, seguindo modelos de propagação da ativação de idéias ("*spreading activation models*") [4]. Esta é uma técnica usada para modelar como a memória humana funciona e assume que a mente associa idéias de acordo com leis psicológicas: um leve odor de perfume pode fazer alguém se lembrar da namorada ou da ocasião em que ela usou o perfume. Algum sinal (no caso, o perfume), ativa uma idéia dormente no cérebro, e essa idéia ativa outra, sucessivamente.

Segundo Pirolli [25], a web é formada por inúmeras "regiões" ou "campos de informação" (por exemplo, seções de *sites* ou a lista de resultados de uma busca) onde a "distância" entre dois itens de conteúdo da mesma região é menor do que entre itens de regiões diferentes. Por esses estudos, é mais fácil navegar e processar informação que reside dentro de uma mesma "região" e, portanto, projetar uma organização que aproxime e agrupe conteúdos relacionados tenderia a facilitar ao usuário encontrá-los.

Enquanto essas táticas referem-se a organizações criadas e mantidas pelos produtores do *site*, por outro lado, a *folksonomia* propicia a criação de espaços colaborativos onde as pessoas podem gerar novas opções para organizar, apresentar, navegar e recuperar informações. *Folksonomia*[1], que pertence a um conjunto de abordagens da chamada Web 2.0 [24], é uma maneira de indexar informações usando uma rotulação aberta baseada em "*tags*" (etiquetas). Na elaboração de uma taxonomia convencional, em geral o

[1] http://vanderwal.net/folksonomy.html

primeiro passo é definir as categorias para depois associar os itens de conteúdo, enquanto na *folksonomia* cada usuário analisa o item de conteúdo para depois classificá-lo com "*tags*".

A capacidade de se recuperar conhecimento de comunidades pode trazer grandes benefícios, como identificar correlações entre os conteúdos com base nos interesses dos usuários e não nas suposições do arquiteto de informação. Mas aproveitar as classificações geradas pelos usuários nessa abordagem, com o objetivo de facilitar as tarefas de pesquisa exploratória, ainda é um desafio, pois em geral, o conhecimento gerado não é estruturado e há problemas de falta de consistência [23, 24]. Iniciativas recentes objetivam prover uma abordagem semântica às iniciativas da Web 2.0 através da criação de ontologias colaborativas [23]. Porém, a contribuição dos usuários ocorre durante a administração do *site*, ou seja, *a posteriori* da implantação, quando, e quanto mais os usuários participam. Portanto, seria útil uma forma de aproveitar a colaboração dos usuários ainda na fase de projeto do *site*.

Card Sorting e Contexto na Pesquisa Exploratória

No projeto do sistema de organização, o *card sorting* é uma técnica bastante sugerida por vários autores para auxiliar na tarefa de classificar e organizar o conteúdo de forma centrada no usuário [21, 26], e que auxilia a identificar as diferenças de perspectivas existentes entre os diferentes grupos de usuários.

Para classificar itens de conteúdo textuais (por exemplo, artigos sobre determinado assunto) em um experimento de *card sorting*, normalmente os participantes dispõem apenas do título (ou pequenos resumos) como única informação perceptível sobre cada item. Procura-se dar ao usuário a mesma informação que ele teria ao navegar pelo *site*. A partir desses itens, que servem como indexadores de conteúdo, a forma que cada usuário escolhe para agrupá-los é baseada na semântica, convenções, subjetivismo e preferências pessoais. Dado um número ***n*** de pessoas, podemos ter ***n*** classificações diferentes, mas os algoritmos de cálculo de distância do *card sorting* propõem-se a atingir uma aproximação do que seria uma organização consensual, embora, em geral, não haja uma organização única que seja ideal para todos os usuários.

Segundo alguns autores [17], como essa técnica é bastante centrada no conteúdo, é possível que a forma de classificar o conteúdo não esteja relacionada com o modo como os usuários usarão essa estrutura em uma tarefa de pesquisa exploratória; por isso, um resultado de *card-sorting* pode não ser adequado para esse tipo de pesquisa. Isso acontece porque o título pode não fornecer uma correta visão sobre o que trata determinado artigo. Ao ler, o leitor vai formando uma idéia tanto sobre o tema quanto sobre o **contexto** ao qual o texto se aplica. Uma das definições para "contexto" é o de "inter-relação de circunstâncias que acompanham um fato ou uma situação" [9]. Portanto, pode-se dizer que cada texto está associado a algum contexto.

Em um texto, há uma inter-relação entre os termos usados e o contexto. Os termos em um documento ajudam a definir um contexto, porém, como um termo pode carregar ambiguidades, é o contexto que o especifica (o significado de um termo em uma área do conhecimento é aquele adotado pelos autores dessa área) [6]. No caso da busca exploratória, é o contexto que determina se um texto é ou não útil. Textos podem usar termos distintos, mas estar associados a um mesmo contexto. Um texto pode usar o termo "passeio" e, outro, o termo "jogos", mas ambos poderiam tratar do tema "lazer com os filhos". A análise do leitor sobre o conjunto e relacionamento dos termos nos textos pode comprovar, ou não, cada caso.

Dessa forma, mesmo no caso de textos que não possuam similaridade semântica (isto é, não usem das mesmas palavras-chave), ou tratem de assuntos diferentes, por exemplo, um sobre "amamentação", e o outro sobre "fraldas", uma pessoa pode considerar que ambos os textos possuem um mesmo contexto, ou que possuem, ao menos, o que se poderia chamar de um contexto aproximado.

Entretanto, estabelecer que dois textos possuam uma proximidade de contexto, e avaliar esse grau de proximidade, em sendo possível estabelecer esse tipo de medida, envolve certa dose de subjetividade, pois depende, por exemplo, do nível de conhecimento do avaliador em relação à área do conhecimento no qual os textos estão inseridos. Mesmo que essa avaliação seja feita por algum especialista nos assuntos tratados pelos textos, sempre se pode questionar se essa associação é de senso comum para a maioria das pessoas, e, se não for, qual a avaliação que deve ser adotada em um projeto centrado no usuário.

Conhecimento de Senso Comum

Alguns projetos, como o OMCS[2] (*Open Mind Common Sense* - MIT) e sua versão brasileira, o OMCS-Br[3] (UFSCar – Universidade Federal de São Carlos) coletam conhecimentos que compõem o que se denomina senso comum das pessoas. A comunidade de Inteligência Artificial usa esse termo para se referir aos fatos básicos e entendimentos que a maioria das pessoas tem [16], como, por exemplo, saber que um quarto de hotel deve ter uma cama, pessoas possuem unhas, gelo é frio, Coca-Cola é um refrigerante e um filho é mais jovem que seu pai.

Sites e jogos interativos têm sido utilizados para motivar colaboradores a contribuírem na coleta de senso comum, possibilitando a aquisição de grande volume de conhecimento sobre diversos temas. O que os usuários consideram como correto e normal, e o que não parece fazer sentido, é um dos resultados que podem ser obtidos

[2] http://openmind.media.mit.edu/

[3] http://www.sensocomum.ufscar.br

dessa coleta, definindo-se senso comum como o conhecimento compartilhado pela maioria das pessoas de uma determinada cultura [5]. Dizer que uma afirmação é senso comum em uma cultura não significa que seja cientificamente verdadeira ou que seja senso comum em outras culturas [1], mas mesmo esse tipo de afirmação faz parte do escopo de coleta desses projetos.

Uma das abordagens para representação do conhecimento coletado é a utilização de redes semânticas (denominada ConceptNet no projeto OMCS, e ConceptNetBr, no OMCS-Br) que relacionam conceitos através de relações binárias semânticas (ilustradas na Figura 1) que abrangem os vários tipos de conhecimentos (por exemplo: propriedades; causa e efeito; generalização/especialização) representados por predicados das relações do tipo "LocationOf", "PropertyOf", "UsedFor" e outros.

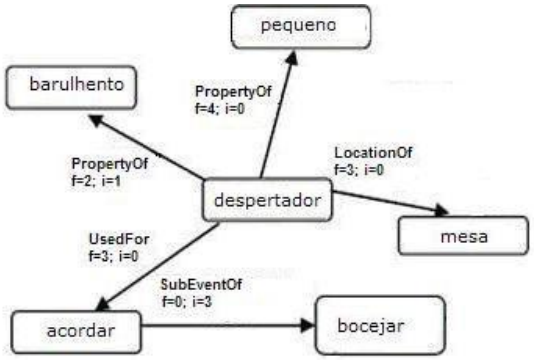

Figura 1 - Rede de conceitos ConceptNetBr [5]

Cada tipo de relação possui uma carga semântica como, por exemplo, relações do tipo "LocationOf X Y" (LocationOf "Los Angeles" "Califórnia), carregam consigo a semântica de que X pode ser encontrado em Y [5].

O conhecimento das pessoas é coletado e inicialmente armazenado na base de conhecimento na forma de sentenças, formando uma ontologia baseada em padrões de sentenças, e a representação do conhecimento em linguagem natural permite coletar modos ambíguos e redundantes de expressar uma mesma idéia. Por exemplo, "carro" e "automóvel" expressam quase o mesmo, mas pode implicar em diferenças contextuais, como o aspecto de formalismo em um discurso [16].

No processo de extração e análise dos dados da base de conhecimento para geração da rede semântica, a frequência (f) com que sentenças fornecidas pelos colaboradores dão origem a uma relação entre dois termos é contabilizada, presumindo-se que quanto mais pessoas forneçam as mesmas sentenças, maior é a probabilidade de que a relação represente um conhecimento de senso comum.

Algumas relações podem ser geradas automaticamente por regras de inferência. Por exemplo, uma vez encontradas as relações (IsA "maçã" "fruta") e (IsA "banana" "fruta"), e as relações (PropertyOf "maçã" "doce") e (PropertyOf "banana", "doce"), por silogismo pode-se gerar a relação (PropertyOf "fruta" "doce") [5].

Já existem aplicações que se utilizam de bases de conhecimento de senso comum, como o projeto KitchenSense [15], uma aplicação integrada por meio de sensores a uma cozinha e a seus utensílios, em que o sistema infere o relacionamento causal e temporal de eventos para detectar o comportamento dos usuários e prever suas intenções dentro do contexto para, com isso, oferecer ajuda e sugestões nos procedimentos e simplificar as interfaces de controle de alguns dispositivos de acordo com sua relevância para a tarefa que o usuário pretende fazer, tornando o ambiente mais "inteligente". Outras aplicações usam esse tipo de base de conhecimento para, por exemplo, atuar como tradutores de textos que levam em consideração diferenças culturais [2].

Essas aplicações mostram que a partir do conhecimento do que é de senso comum para a maioria das pessoas pode-se desenvolver sistemas que identifiquem relações entre conteúdos, que façam sentido para as pessoas. A rede semântica relaciona os conceitos usados pelas pessoas, tornando possível identificar quais os termos existentes em um dado documento são mais significativos, por terem mais conceitos relacionados, e também extrair os conceitos mais relevantes relacionados a um determinado termo, domínio de aplicação ou tema. É possível, dessa forma, obter um contexto de relacionamento entre conceitos que fazem parte do senso comum das pessoas, e também identificar novos conceitos e correlações que podem não ser do conhecimento do arquiteto de informação.

ORGANIZAÇÃO COM BASE NO CONTEXTO DE SENSO COMUM

Foi visto que o conhecimento de senso comum pode ser coletado, armazenado e relacionado através de uma rede de conceitos, como na ConceptNetBr do projeto OMCS-Br. Usamos essa base em nosso trabalho anterior [28], cujos resultados são sumarizados neste artigo, mostrando que:

a) Em alguns casos, para um dado item de conteúdo textual, é possível identificar, em uma base de conhecimento de senso comum, o conjunto de conceitos relacionados a esse conteúdo, usado para compor uma expressão de natureza semântica capaz de representar não apenas o conteúdo de um texto, mas também seu contexto.

b) Uma vez obtido o contexto de senso comum de cada item de conteúdo, é possível gerar uma medida que avalie o "grau de similaridade de contexto" para identificar os itens de conteúdo próximos.

c) Calculado o "grau de similaridade de contexto" entre os itens de conteúdo de um *site*, é possível gerar uma organização para esse *site* que denominamos "Esquema de Organização com base no Contexto de Senso Comum" ou "CSCOS – *Common Sense Context Organization Scheme*", porque o contexto associado ao conteúdo é atribuído pelo senso comum das pessoas.

A premissa, portanto, para que esse processo seja usado, é o da existência do objetivo de se projetar um *site*, e da pré-existência de itens de conteúdo que deverão fazer parte desse *site*, os quais precisam ser organizados. Sumarizamos, a seguir, a estratégia utilizada no CSCOS para obter uma organização usando uma base de conhecimento de senso comum.

Processo de Geração

Como um *site* pode ser constituído de um grande volume de conteúdo, foi desenvolvido um software de apoio (**CSCOSGenerator,** composto de diversos módulos), para automatizar o processo proposto (Figura 2).

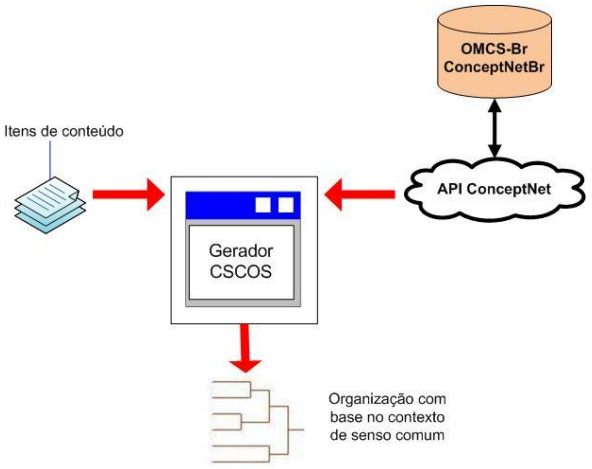

Figura 2 – Processo de geração CSCOS [28]

O processo proposto contempla as seguintes atividades:

A1 – Selecionar o acervo de itens de conteúdo (artigos na forma de textos) que serão analisados e organizados, gerando uma lista, basicamente contendo o título de cada artigo, um número sequencial (ID) de identificação, e o endereço do arquivo físico que contém o texto digitalizado.

A2 – Identificar as principais palavras significativas existentes em cada item de conteúdo (palavras-chave), que caracterizam o teor de cada texto, formando o que se denominou sua "assinatura léxica" (AL), e a frequência com que cada palavra-chave é utilizada, objetivando caracterizar sua importância no texto. Palavras consideradas irrelevantes (as "*stop-words*", por exemplo, conjunções, artigos, preposições etc.) foram descartadas. Após a extração (usando técnicas de "*Text Mining*" e "*tokenização*"), os termos extraídos passam por um processo de normalização ("*stemming*"), buscando-se sua forma canônica, como, por exemplo, substantivos e adjetivos são postos no singular, e verbos postos no infinitivo. Evita-se, assim, que termos sejam considerados distintos apenas em função de sua variação morfológica. Dessa forma, a assinatura léxica (AL) de cada item de conteúdo i é formada por um conjunto dos termos significativos (Tx) mais usados no texto, cada qual com sua frequência de uso (fx): $AL_i = \{T1_{(f1)}, T2_{(f2)}, ...Tn_{(fn)}\}$.

A3 – Buscar, na ConceptNetBr, os conceitos que o senso comum relaciona a cada palavra-chave, o tipo de relação semântica e a frequência com que essas relações ocorrem. Consideramos que esses conceitos seriam similares às palavras-gatilho que o senso comum das pessoas associa às palavras-chave do texto, e que cada palavra-chave seria uma "semente de pesquisa" na rede semântica. Para isso foi construído um componente de software que se utiliza de uma API (*Application Programming Interface*) da ConceptNetBr para realizar consultas on-line via Internet. Tendo em vista que diferentes palavras-chave podem trazer conceitos de senso comum (termos) em repetição, esses conceitos são agrupados selecionando-se os mais relevantes (de maior frequência). Esse conjunto de conceitos de senso comum foi denominado "contexto de senso comum" do texto.

Considerou-se que, através dessa atividade, foi possível descobrir qual o contexto de que trata determinado item de conteúdo, sem se limitar aos significados das palavras usadas no texto. A palavra-chave "copo", por exemplo, poderia estar associada no senso comum aos conceitos de "vidro" ("MadeOf") e "cozinha" ("LocationOf") que talvez tenham pouca relação com o texto. Mas, considerando-se a frequência das associações em todo o conjunto de conceitos associados às palavras-chave do texto, o processo propõe identificar o contexto de forma automática, com os conceitos que as pessoas mais associam às palavras usadas no texto, sem a necessidade de um especialista ou pessoa para ler e interpretar o texto. Por exemplo, pode-se descobrir que as palavras "bicicleta" e "dominó" são frequentemente associadas aos conceitos de "criança" e "lazer" no senso comum das pessoas.

A4 – Identificados os termos que representam os conceitos de senso comum para cada item de conteúdo, esses itens foram classificados conforme a similaridade entre seus contextos, gerando o que se denominou um esquema de organização com base no contexto de senso comum. Essa parte do processo é decomposta nas seguintes etapas:

A41 – Nesta etapa, o objetivo é consolidar o conjunto de termos que caracterizam o contexto de senso comum (TC) de cada item de conteúdo. Para cada item de conteúdo foi gerado o que se denominou "assinatura léxico-contextual" (ALC), contendo os termos do contexto de senso comum e as palavras-chave do texto (TC + ALC), visto que as palavras-chave caracterizam o respectivo texto.

Tanto os termos da "assinatura léxica" quanto os da "assinatura contextual" possuem associado um valor de frequência referente ao seu peso no respectivo conjunto, com significados diferentes: no primeiro caso, a quantidade de vezes que o termo foi usado no texto e, no segundo, a frequência das relações semânticas do termo na rede de conceitos de senso comum, decorrente da contribuição dos colaboradores do projeto OMCS-Br ou por regras de inferência. Por isso, antes de gerar a "assinatura léxico-contextual", esses valores foram normalizados usando a

técnica TF-IDF, muito citada em algoritmos de recuperação da informação e de comparação de similaridade entre documentos [10]. A técnica atribui um peso a cada termo, conforme o número de ocorrências desse termo no documento (TF - "*Term Frequency*") e em razão de sua importância no conjunto de documentos da coleção (IDF - *"Inverse Document Frequency"*). No primeiro caso, considerou-se como coleção o conjunto das "assinaturas-léxicas" (AL), e no segundo, o conjunto dos termos de contexto (TC).

A42 – Nesta etapa o objetivo é comparar as assinaturas léxico-contextuais dos itens de conteúdo, dois a dois, e calcular o grau de similaridade entre eles. Denominou-se "similaridade contextual" uma medida que expressa o quanto dois itens tratam dos mesmos assuntos em um mesmo contexto, considerando, para esse cálculo, a quantidade e peso dos termos que existem em comum em suas assinaturas. O produto dessa etapa será uma matriz de similaridade **n x n**, que mostra o grau de similaridade entre cada par de itens de conteúdo (Figura 3).

ID	1	2	3	4	5
1	0	0,259811	0,025149	0,022868	0,369446
2	0,259811	0	0,077223	0,006029	0,368352
3	0,025149	0,077223	0	0,026535	0,023596
4	0,022868	0,006029	0,026535	0	0,016377
5	0,369446	0,368352	0,023596	0,016377	0
6	0,030936	0,052546	0,027477	0,010758	0,029425
7	0,017964	0,039905	0,031312	0,006258	0,024716
8	0,030915	0,077746	0,056562	0,02632	0,021987
9	0,027776	0,039525	0,072941	0,015695	0,030048
10	0,018041	0,047592	0,037158	0,005797	0,017796
11	0,064369	0,031909	0,01504	0,032951	0,002273
12	0,064376	0,114264	0,043796	0,024551	0,011729

Figura 3 – Trecho da matriz de similaridade [28]

Essa conceituação simplificada de "similaridade" considera que as assinaturas léxico-contextuais contêm e expandem os termos do léxico usando o senso comum (uma expansão que possui semelhança a algo que talvez só fosse possível obter com um *thesaurus*; e uma aproximação do conceito mais extenso de "similaridade" proposto por Gentner [12], que distingue quatro tipos de similaridades: *"analogy", "literal similarity", "relational abstraction" e "mere-appearance match"*).

Diversas técnicas usadas para estimar a similaridade entre documentos são baseadas na co-ocorrência de termos em *corpus* de documentos [4]. Optamos pela Medida de Similaridade por Cosseno, uma técnica largamente utilizada [3], usando como frequência dos termos o valor TF-IDF calculado anteriormente.

A43 – Em seguida, a matriz de similaridade é submetida a um algoritmo de análise de agrupamento ("*cluster analysis*"), para compor uma hierarquia de grupos de classificação. Na abordagem denominada *clustering* hierárquico, os agrupamentos são hierarquizados, sendo que cada grupo pode ser subdividido em grupos menores, e assim sucessivamente [18]. Para isso foi utilizada a ferramenta statistiXL[4], que funciona como um complemento (*add-in*) da planilha Microsoft Excel.

Com base nos dados dessa matriz, a análise de agrupamento gera um diagrama hierárquico (dendograma), produto final do processo (Figura 4), que sugere a estrutura básica de um sistema de organização, da mesma forma que o resultado obtido pela técnica de *Card Sorting*.

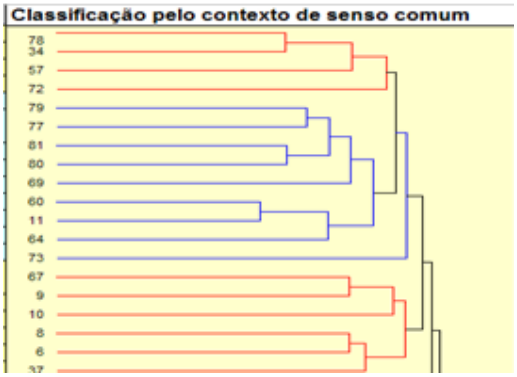

Figura 4 – Trecho do dendograma gerado [28]

O dendograma apresenta uma hierarquização de grupos e sub-grupos que pode ser interpretada de várias maneiras, pois há grupos com muitos níveis hierárquicos, e outros com poucos níveis mas muitos sub-grupos em cada nível. Segundo Garrett [11], dados de pesquisa e metodologias formalizadas não garante, por si só, uma arquitetura de informação melhor, e, por isso, qualquer método que não aborde o processo criativo estará lamentavelmente incompleto. Ou seja, é necessário uma análise crítica e um refinamento por parte do arquiteto da informação, sendo razoável supor que, quanto mais precisos os dados da pesquisa, mais facilitado fica esse trabalho, e nesse ponto, o CSCOS pode ser muito útil.

A seguir, como prova de conceito, apresentamos um estudo de caso comparativo que utiliza esse processo e compara os resultados obtidos com um estudo de *card sorting*. E na sequência, nossas conclusões e discussões finais.

PROVA DE CONCEITO

Nesta seção, descreve-se um experimento que utilizou o processo proposto neste trabalho (CSCOS) para obter um esquema de organização para um *site*, utilizando-se da Base de Conhecimentos de Senso Comum do projeto OMCS-Br (*Open Mind Common Sense* no Brasil).

Foram escolhidos 81 textos para compor um acervo que versa sobre o tema Pais & Filhos (por exemplo, artigos sobre educação de filhos, desde a infância até a adolescência, educação sexual, sinopses de livros sobre o tema, lazer com os filhos, e outros) que deveriam compor um novo *site*. Esses textos correspondem a uma amostra da seção Pais & Filhos de um *site* pré-existente (Portal da

[4] http://www.statistixl.com

Família[5]). Embora o foco do nosso processo seja para a **etapa de projeto** de *sites*, utilizamos textos prontos, reais, para fins de agilizar o presente estudo. Optou-se pelo tema "pais e educação dos filhos", principalmente porque o projeto OMCS-Br já possui dados relacionados a esse tema (entre outros) e sobre atividades afins.

Também usamos o mesmo conjunto de textos em um experimento de *card sorting*, que gerou um esquema de organização, para o mesmo *site*, com o objetivo de possibilitar uma análise comparativa entre os dois esquemas de organização. Para o experimento, utilizou-se uma ferramenta de *card sorting online* (*WebSort*[6]), com a participação de 27 voluntários. Cada participante gastou cerca de 20 a 30 minutos no experimento, tendo apenas o título do texto como referência.

A Figura 5 representa o esquema de organização sugerido pelo processo CSCOS. Os números entre parênteses representam a quantidade de textos em cada grupo.

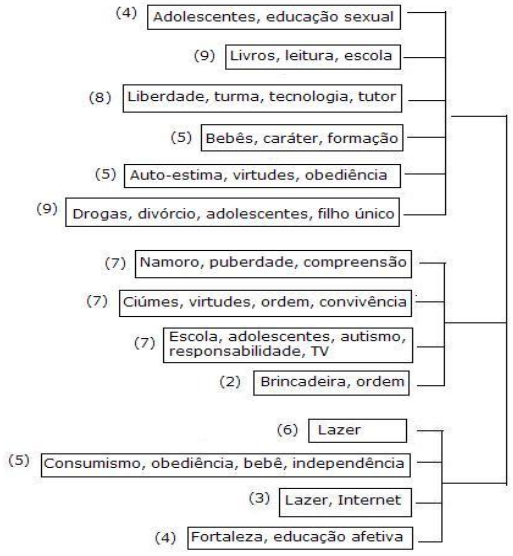

Figura 5 – Organização sugerida pelo CSCOS [28]

Com o uso dos softwares de apoio o processo de geração CSCOS demorou menos de 15 minutos para ser concluído. Cada texto do acervo usado possuía, em média, 164 palavras-chave distintas (mínimo de 53, máximo de 791) variando conforme o tamanho do texto, a abrangência de assuntos e a variedade do vocabulário usado pelo autor. Observou-se também que algumas palavras-chave são muito utilizadas por alguns textos (por exemplo, a palavra "criança" é usada 163 vezes em um dos textos).

Optou-se por uniformizar o tamanho da assinatura léxica (AL) com as 40 palavras-chave mais frequentes de cada texto, que corresponde a 25% da quantidade média de palavras-chave por artigo do acervo (embora corresponda a 75% das palavras-chaves do menor artigo, que possui 53 palavras-chaves). O tamanho da assinatura léxica ainda é uma questão em aberto, e deve levar em consideração o tamanho dos textos do acervo e o retorno da busca que será feita na ConceptNetBr com cada palavra-chave.

Houve um total de 1.244 palavras-chave únicas (isto é, sem repetição) no conjunto das assinaturas léxicas (AL). Dessas, 72% foram encontradas na *ConceptNetBr* e retornaram relações de contexto. Consideramos que o fato de nem todas as palavras-chave encontrarem conceitos correspondentes na base de senso comum não representa um problema grave neste trabalho de investigação, tendo em vista que a assinatura léxico-contextual (ALC) será composta, na estratégia adotada, tanto por palavras-chave quanto por termos oriundos do contexto de senso comum.

Cada assinatura léxica recuperou, em média, 3.915 termos de contexto do senso comum da ConceptNetBr, sem repetição. O tamanho da assinatura léxico-contextual (ALC) foi uniformizado em 80 termos, sendo que a determinação deste número é ainda uma questão não resolvida. Trata-se de uma quantidade pequena, em relação ao conjunto de termos associados por texto, mas considerável se fossem "*tags*" a ser atribuídas por pessoas ainda na fase de projeto do *site*.

O esquema de organização gerado pelo processo CSCOS faz sentido do ponto de vista matemático (quantidade e frequência dos termos em comum). Por isso, para uma melhor análise, realizou-se uma comparação com o esquema obtido pela técnica de *card sorting* (Figura 6).

A comparação analisou detalhes dos agrupamentos de itens, hierarquização e relacionamento dos agrupamentos. Foram encontrados tanto casos de semelhanças quanto de divergências (por *card sorting* e por similaridade contextual), apresentados aqui de forma resumida. Por exemplo, os textos nº. 31 ("Dicas de passeios e lazer") e nº. 48 ("Jogos e brincadeiras de salão") foram agrupados tanto no *card sorting* quanto no CSCOS, apesar de terem apenas três termos em comum em suas assinaturas léxicas (7,5% do total de 40 termos). Ou seja, os participantes do *card sorting* consideraram que, mesmo tratando de situações concretas diferentes, há uma alta proximidade entre esses dois textos, apenas analisando seus títulos. De fato, verificou-se que há 23 termos em comum nas suas assinaturas léxico-contextuais (28,7% de um total de 80 termos), evidenciando a similaridade contextual.

[5] http://www.portaldafamilia.org

[6] http://www.websort.net/

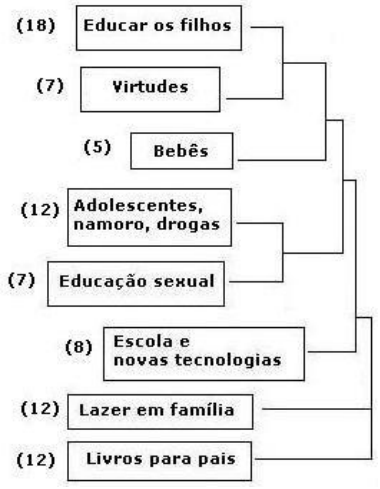

Figura 6 – Dendograma resumido resultante do *card sorting* [28]

Algumas diferenças interessantes foram observadas. No *card sorting* os textos referentes a sinopses de livros (todos possuíam a palavra "livro" nos seus titulos) foram agrupados, enquanto que a organização por contexto de senso comum espalhou esses textos por outros grupos. No primeiro caso, esse tipo de agrupamento parece facilitar a pesquisa por um tipo de item específico (por exemplo, um livro sobre como cuidar de bebês), enquanto que, no segundo, parece ser o mais apropriado para pesquisas exploratórias (por exemplo, sobre cuidados com bebês, sejam em forma de artigos, livros, filmes ou outros).

Na organização por *card sorting*, os textos nº 50 ("Meu filho é um irresponsável") e nº 49 ("Meu filho é desobediente?") foram agrupados. De fato, a similaridade entre os títulos dos textos parece evidente (ambos são direcionados aos pais e tratam de comportamentos dos filhos). Mas, na organização por contexto de senso comum, o texto nº. 50 apresentou maior similaridade com os textos nº. 61 ("Os pais perante o rendimento escolar") e nº. 19 ("Boas notas na escola"), e o texto nº. 49 maior similaridade com o texto nº. 74 ("Livro: Cuidando do bebê até que sua mulher chegue em casa"). Comparando-se suas assinaturas (ALC), verificou-se que os textos nº. 50 e nº. 49 possuem apenas 31 termos em comum, enquanto que o texto nº. 50 possui 41 termos em comum ao nº. 61, e 37 em relação ao texto nº. 19.

O fato de o texto nº. 49 ter sido agrupado ao texto nº. 74, por similaridade contextual, encontra uma explicação pelo fato de o texto tratar da obediência que, segundo David Issacs[7], um especialista mundial no assunto, é uma virtude a ser trabalhada pelos pais principalmente em crianças de zero a 7 anos, tendo em conta os traços estruturais das idades e a natureza das virtudes, visto que "1) se não se desenvolve desde pequenos, é muito mais difícil depois; 2) é uma virtude necessária para uma convivência feliz; 3) tranquiliza as mães". Já a virtude da responsabilidade, tratada no texto nº. 50, é tratada com maior ênfase desde os 8 até os 12 anos (ou seja, na idade escolar) pois necessita de maior uso da vontade "para suportar incômodos, esforçar-se continuamente, alcançar o decidido e resistir a influências nocivas".

Nesse exemplo, a classificação por senso comum detectou casos de similaridade que só seriam estabelecidos com uma análise mais detalhada de cada texto, realizada por um especialista no tema, e que não haviam sido evidenciadas na classificação gerada através do *card sorting*. Também evidencia casos que se poderia denominar de "falsos similares", ou seja, em que os usuários agrupam dois textos com baixo grau de similaridade contextual.

Entretanto, um teste que permita comparar qual das organizações (por CSCOS e por *card sorting)* é melhor em determinada situação não é simples. Técnicas propostas na literatura, usadas para pré-validar uma organização na fase de projeto, como, por exemplo, construir protótipos do *site* que implementam a estrutura de navegação, mas despojados de outros aspectos gráficos e testá-los com usuários [22], não se adequam nesse caso, justamente porque a identificação da similaridade de textos nem sempre é algo intuitivo somente com a observação dos títulos dos textos. E, apenas com os títulos, a organização proposta pelos usuários é, em tese, a resultante do *card sorting*. Além disso, a organização com base na similaridade de contexto seria, como explanado neste trabalho, mais adequada a usuários durante uma pesquisa exploratória sobre casos concretos, onde fatores como emoções, incertezas e pressão de tempo afetam a qualidade e a natureza da pesquisa e dos seus resultados. Portanto, não se considera uma questão trivial conseguir colaboradores representativos dos grupos de usuários e compor cenários de tarefas que possibilitem um teste entre os dois tipos de organização, e, dessa forma, este trabalho não conduziu esse tipo de comparação.

Adicionalmente, geramos uma organização por similaridade usando apenas as palavras-chave dos textos, obtendo uma classificação que também teve pontos em comum quanto pontos divergentes, tanto em relação ao gerado pelo CSCOS quanto ao gerado por *card sorting*.

Em resumo, consideramos que o processo proposto funcionou a contento em nossa prova de conceito, gerando um esquema de organização com coerência nos agrupamentos propostos conforme a similaridade de contexto, alternativo em relação ao que seria gerado por *card sorting*, e em um tempo muito menor do que este, quando se faz uso de artefatos de software que automatizem os cálculos manipulando uma quantidade considerável de termos.

CONCLUSÕES E CONSIDERAÇÕES

Segundo nossa avaliação, os resultados obtidos permitiram concluir a viabilidade do processo proposto para geração do CSCOS, o esquema de organização com base no contexto

[7] Esquema de virtudes por idades - www.portaldafamilia.org/artigos/virtesq1.shtml

de senso comum. A vantagem de se usar a similaridade de contexto, segundo nosso ponto de vista, é não restringir a avaliação da proximidade de itens à similaridade semântica entre eles.

Nosso processo obtém um esquema de organização que pode ser tanto alternativo quanto complementar ao que seria gerado através da técnica tradicional de "*card sorting*", auxiliando o projetista na identificação antecipada de conteúdos de interesse para usuários para gerar uma organização em tese mais eficiente, que melhor incentivaria o usuário a usar e participar do *site*.

Embora não tenhamos definido uma estratégia capaz de avaliar se alguma das classificações é melhor que a outra (*card sorting* x similaridade contextual), evidenciou-se que o CSCOS auxilia o projetista a identificar relações de similaridade que podem não ter sido percebidas em um estudo de *card sorting*, e a identificar falsas similaridades. Porém, no processo CSCOS, além de não se ter uma lista de "sugestões de rótulos" para cada agrupamento (um subproduto gerado no *card sorting*), pode ocorrer certa dificuldade para criar rótulos curtos (ou seja, com poucas palavras) que represente o contexto de um grupo de itens.

Os resultados apresentados permitem concluir que a participação das pessoas no desenvolvimento do esquema de organização de *sites*, ainda que de forma indireta, pode ser obtida através de Bases de Conhecimento de Senso Comum, aproveitando-se do conhecimento social das pessoas que participaram e colaboraram para relacionar conceitos na coleta de senso comum. Nesse sentido, trata-se de um contraponto em relação aos benefícios adquiridos *a posteriori* nas abordagens mais comuns da Web Social.

Há, porém, diversas questões em aberto no processo. Por exemplo, os testes realizados sobre o impacto das variações das quantidades de termos nas assinaturas léxica e léxico-contextual não foram conclusivos, e por isso, essa questão foi deixada em aberto para trabalhos futuros, bem como validar a aplicabilidade do processo em grandes acervos.

Deve-se ter em consideração que nem todo tipo de conhecimento pode ser considerado como sendo de senso comum. Conteúdo direcionado a assuntos técnico-científicos, por exemplo, ou da intranet corporativa de uma empresa, podem não encontrar respaldo em uma base de conhecimento de senso comum. Porém, como o projeto OMCS-Br possibilita coletar conhecimento de senso comum sobre diversos temas, há muitos tipos de projetos onde se poderia aplicar o processo proposto (por exemplo, *sites* de notícias, educação, lazer e assuntos do cotidiano).

Ao apresentar uma proposta de processo de produção da organização com base no senso comum, este trabalho abre o ensejo para novos estudos que inovem ou melhorem esse processo. Trabalhos futuros poderiam explorar outras potencialidades, como a identificação automática de contexto de forma facetada. Se existirem palavras-chave que descrevam as perspectivas (facetas) de cada categoria, seria possível obter uma assinatura léxico-contextual para cada uma dessas perspectivas. Um novo item de conteúdo, publicado por um participante em um site social, por exemplo, poderia ser avaliado conforme sua similaridade de contexto em relação às perspectivas existentes, e automaticamente classificado pelo sistema de publicação.

Ou um sistema de recomendação por contexto de senso comum, uma abordagem com foco no conteúdo condizente com a recomendada por alguns autores [13, 27]. Ao visitar uma página, o usuário teria à disposição uma lista de sugestões de páginas contextualmente similares à atual. Em nosso trabalho anterior [28] sugerimos apresentar a assinatura (ALC) de cada item da lista de sugestões na forma de "*text tag clouds*", para aumentar o cheiro da informação e auxiliar o usuário em sua navegação.

AGRADECIMENTOS

Nossos agradecemos especiais à Profª. Dra. Junia Coutinho Anacleto, coordenadora do Laboratório de Interação Avançada (LIA) da UFSCar, e a Fabiano Pinatti, que nos possibilitaram o acesso ao projeto OMCSBr.

REFERÊNCIAS

[1] ANACLETO, J. C.; CARVALHO, A. F. P. DE; PEREIRA, E. N.; FERREIRA, A. M.; CARLOS, A. J. F. Machines with good sense: How can computers become capable of sensible reasoning?. In: BRAMER, M. (Org.). Artif. Intelligence in Theory and Practice II WCC2008. Berlin:Springer-Verlag,2008,v.1,p.1-10.

[2] ANACLETO, J. C.; LIEBERMAN, H.; TSUTSUMI, M.; NERIS, V.; CARVALHO, A. F. P.; ESPINOSA, J.; GODOI, M.; ZEM-MASCARENHAS, S. H. Can Common Sense uncover cultural differences in computer applications? In: Intl. Federation for Information Processing (IFIP), v. 217, 2006. 10 p.

[3] BAYARDO, R. J.; MA, Y.; SRIKANT, R. Scaling up all pairs similarity search. In: International Conference on World Wide Web, 16., Banff, Alberta, Canada, May 08 - 12, 2007. Proceedings... WWW '07. ACM, New York, NY, 131-140.

[4] BUDIU, R.; ROYER, C.; PIROLLI, P. Modeling Information Scent: A Comparison of LSA, PMI, and GLSA Similarity Measures on Common Tests and Corpora. Palo Alto Research Center, CA, 2007.

[5] CARVALHO, A. F. P. Utilização de Conhecimento de Senso Comum no Planejamento de Ações de Aprendizagem Apoiado por Computador. São Carlos, 2007. 257 f. Dissertação (Mestrado em Ciência da Computação) – Univ. Federal de São Carlos, 2007.

[6] CHEN, H.; NG, T. D.; MARTINEZ, J.; SCHATZ, B. R. A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the Worm Community System. Journal of the American Society for Information Science. V. 48 Issue 1, Pages 17 - 31, 1997.

[7] CHI, E. H.; PIROLLI, P. L.; CHEN, K.; PITKOW, J. E. Using information scent to model user information needs and actions and the Web. In: Conference on Human Factors and Computing Systems (CHI 2001); Seattle, WA. NY. Proceedings. ACM; 2001; 490-497.

[8] DERVIN, B.; FOREMAN-WERNET, L. Sense-Making methodology reader. Selected writings of Brenda Dervin. Hampton Press, Cresskill, NJ. 2003

[9] DICIONÁRIO HOUAISS, 2001

[10] FANG, H.; TAO, T.; ZHAI, C. 2004. A formal study of information retrieval heuristics. In: ACM SIGIR Conference on Research and Development in information Retrieval, 27., 2004, Sheffield, United Kingdom. Proceedings...SIGIR '04, ACM.

[11] GARRETT, J. J. ia/recon. Adaptive Path, 2002. Disponível em: <http://www.jjg.net/ia/recon/>.

[12] GENTNER, D. 1988. Metaphors as Structure-Mapping. The Relational Shift. Child Develop. 47-59.

[13] HALLAND, A. G. U., HALLAND, M. S. Core+Paths: A Design Framework for Findability, Prioritization and Value. Netlife Research Usability Specialists. In: ASIS&T, IA Summit 2007. Disponível em: <http://www.iasummit.org/2007/>.

[14] KUHLTHAU, C. C. Seeking Meaning: A process approach to library and information services. Westport, CT: Libraries Unlimited, 2004.

[15] LEE, C. J.; BONANNI, L.; ESPINOSA, J. H.; LIEBERMAN, H.; SELKER, T. Augmenting kitchen appliances with a shared context using knowledge about daily events. In: Intl. Conf. on Intelligent User Interf., 11, Sydney, Australia, 2006. Proceed. IUI '06.

[16] LIU, H.; SINGH P. ConceptNet: a practical common sense reasoning toolkit. BT Technology Journal, v. 22, n. 4, p. 211-226, 2004.

[17] MAURER, D.; WARFEL, T. Card sorting: a definitive guide. Boxes and Arrows, 2004. Disponível em: <http://www.boxesandarrows.com/archives/card_sorting_a_definitive_guide.php>.

[18] METZ, J; MONARD, M. C. 2006. Estudo e Análise das Diversas Representações e Estruturas de Dados Utilizadas nos Algoritmos de Clustering Hierárquico. Relatórios Técnicos do Inst. de Ciências Matemáticas e de Computação (ICMC), Nº 269, ISSN - 0103-2569, USP Campus São Carlos.

[19] MORVILLE, P. Ambient Findability. 1 ed. Cambridge: O'Reilly, 2005a. 188p.

[20] MORVILLE, P. User Experience Design. Semantic Studios, 2004. Disponível em: <http://semanticstudios.com/publications/semantics/000029.php>.

[21] MORVILLE, P.; ROSENFELD, L. Information Architecture for the World Wide Web: Designing Large-Scale Web Sites. 3. ed. Cambridge: O'Reilly, 2006. 504p.

[22] NAKHIMOVSKY, Y.; SCHUSTERITSCH, R.; RODDEN, K. Scaling the card sort method to over 500 items: Restructuring the Google AdWords Help Center. Google - Georgia Institute of Technology. ACM, CHI 2006.

[23] NOY, N. F.; CHUGH, A.; ALANI, H. The CKC Challenge: Exploring tools for collaborative knowledge construction. IEEE Intelligent Systems, Jan/Feb (Vol. 23, No. 1), 2008.

[24] O'REILLY, T. What is web 2.0 – Design Patterns and Business Models for the Next Generation of Software. O'Reilly, 2005. Disponível em: <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>.

[25] PIROLLI, P. Information Foraging Theory: Adaptive Interaction with Information. New York, NY: Oxford University Press, 2007. 204p.

[26] REIS, G. A. dos. Centrando a Arquitetura de Informação no usuário. São Paulo, 2007. 250 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Comunicações e Artes / USP, 2007.

[27] SPOOL, J.; PERFETTI, C.; BRITTAN, D. Designing for the Scent of Information. User Interface Engineering, 2004. 31 p. Livro digital. Disponível em: <http://www.uie.com/reports/>.

[28] WANG, W. S. Uso de uma Base de Conhecimento de Senso Comum em Projetos de Arquitetura da Informação de Web Sites. 2009. 216f. Dissertação (Mestrado em Engenharia da Computação) – IPT Instituto de Pesquisas Tecnológicas do Estado de São Paulo, 2009. Disponível em: <http://www.slideshare.net/wanderleywang/cscos>