

Modelo de Descoberta de Conhecimentos e Interesses Baseado em Insumos Textuais Eletrônicos: Uma proposta para apoio a Gestão do Capital Humano

Model of Knowledge Discovery and Interests Based on Textual Electronics Inputs: A proposal to support the Human Capital Management

Andréa Sabedra Bordin

Doutoranda do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) na Universidade Federal de Santa Catarina - UFSC.

Mestre em Ciência da Computação pela Universidade Federal de Santa Catarina - UFSC. Especialista em Sistemas de Informação pela Universidade Federal de Santa Catarina – UFSC, Santa Catarina, Brasil.

andreabord@gmail.com

Alexandre Leopoldo Gonçalves¹

Doutor em Engenharia de Produção pela Universidade Federal de Santa Catarina - UFSC.

Mestre em Engenharia de Produção pela Universidade Federal de Santa Catarina - UFSC.

Professor do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento –PPGEGC, Universidade Federal de Santa Catarina – UFSC, Santa Catarina, Brasil.

a.l.goncalves@ufsc.br

Aran Bey Tcholakian Morales

Doutor em Engenharia de Produção pela Universidade Federal de Santa Catarina - UFSC.

Professor colaborador do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento – PPGEGC, Universidade Federal de Santa Catarina – UFSC, Santa Catarina, Brasil.

aran@stela.org.br

Editor Científico: José Edson Lara
Organização Comitê Científico
Double Blind Review pelo SEER/OJS
Recebido em 25.04.2013
Aprovado em 30.08.2013



Este trabalho foi licenciado com uma Licença Creative Commons - Atribuição – Não Comercial 3.0 Brasil

¹Bolsista do CNPq - Brasil

RESUMO

O capital humano de uma organização consiste no conjunto de conhecimentos, habilidades, atitudes e interesses de seus colaboradores. Toda a organização deve possuir mecanismos (processos, ferramentas) para gerir seu capital humano, pois este é seu principal recurso. A identificação do capital humano é um processo importante na gerência deste capital, portanto ela deve ser ágil e fornecer uma visão fidedigna desse recurso. O uso de ferramentas de tecnologia de informação pelos colaboradores durante suas práticas de trabalho gera insumos textuais eletrônicos que explicitam seus conhecimentos e interesses e são recursos passíveis de serem analisados para a identificação de tais elementos. Esse artigo apresenta um modelo semiautomático de descoberta de conhecimentos e interesses baseado na análise desses insumos textuais. O modelo utiliza técnicas de Engenharia do Conhecimento como KDT (*Knowledge Discovery in Text*) e ontologia e objetiva ajudar no processo de identificação do capital humano na organização.

Palavras-chave: capital humano, gestão do conhecimento, engenharia do conhecimento, descoberta de conhecimento em texto, ontologia.

ABSTRACT

The human capital of an organization is the set of knowledge, skills, attitudes and interests of its employees. The entire organization must have mechanisms (processes, tools) to manage its human capital, as this is its main asset. The identification and mapping of human capital is an important process in managing human capital. So he must be agile and provide an updated and reliable information of such capital. The use of various tools of information technology by employees during their working practices generate electronic textual inputs that explain their knowledge and interests, and these resources are suitable to be analyzed to identify of such elements. This paper presents a semiautomatic model to discovery of knowledge and interest based on the analysis of electronic textual inputs. The model uses Knowledge Engineering techniques such as KDT (*Knowledge Discovery in Text*) and ontology and it's mains goal is to help in identifying of human capital in the organization.

Keywords: human capital, knowledge management, knowledge engineering, knowledge discovery in text, ontology.

1 INTRODUÇÃO

Estratégias de competitividade ligadas exclusivamente à produtividade vêm se esgotando e o diferencial competitivo vem se focando mais no intangível, naquilo que é chamado de capital intelectual. Para Stewart (2002), o Capital Intelectual constitui a matéria intelectual - conhecimento, informação, propriedade intelectual, experiência - que pode ser utilizada para gerar riqueza. É uma tríade formada pelo capital humano, capital de relacionamento e capital organizacional, sendo o capital humano caracterizado pelas pessoas, seus conhecimentos, habilidades, atitudes e interesses. Nesse cenário em que o capital intelectual é valorizado, as organizações passam a ser denominadas “organizações de conhecimento”. Nonaka e Takeuchi (1997); Sveiby (1999); Stewart (2002) e Edvinsson e Malone (1998) são autores proeminentes na discussão nesse tema.

Em organizações baseadas em conhecimento há uma forte consciência da necessidade de gestão do capital humano, ou seja, da gestão dos colaboradores dessas organizações, conhecidos como trabalhadores de conhecimento. Segundo Drucker (1999), o crescimento de organizações baseadas em conhecimento depende da identificação e cultivo das competências de seus membros.

A função de gerenciar o capital humano, independente do processo de gestão a ser utilizado, pressupõe que seus elementos constituintes (conhecimentos, habilidades, interesses, etc.) sejam identificados, pois não se pode gerir o que não é conhecido. A partir dessa identificação ou mapeamento, várias ações de gestão podem ser tomadas.

Esse artigo apresenta um modelo para identificar conhecimentos e interesses de colaboradores de uma organização a partir dos insumos gerados pela utilização de ferramentas de tecnologia de informação para execução de atividades no ambiente de trabalho. O modelo é baseado em técnicas de Engenharia do Conhecimento e objetiva fornecer subsídios para apoiar o processo de gestão do capital humano. Para discutir o tema proposto, este artigo está organizado em cinco seções, incluindo esta Introdução. Na seção 2, discute-se o processo de identificação de capital humano nas organizações e seus problemas. Em seguida, na seção 3, são discutidas as novas possibilidades de identificação de

conhecimentos e interesses. Na seção 4, apresenta-se o modelo proposto e o fluxo de aplicação desse modelo dentro de uma organização e na seção 5 são feitas as considerações finais sobre o modelo apresentado.

4.1.2 O PROCESSO DE IDENTIFICAÇÃO DO CAPITAL HUMANO

O processo de identificação do capital humano normalmente é executado por meio de técnicas como entrevistas com colaboradores, grupos focais, observação direta, etc.

Observa-se que a aplicação dessas técnicas, que são essencialmente manuais, é burocrática, demanda muito tempo e quando concluída pode não refletir mais as competências de um ambiente de trabalho dinâmico, fazendo com que o mapeamento final apresente certo nível de obsolescência. Além disso, essas técnicas possuem um fator subjetivo forte, pois dependem de um ator humano, o que reduz a fidedignidade das informações no final do processo.

Alguns autores apontam problemas nesse processo de identificação. Para Sandberg (2000) a identificação prévia do que o colaborador sabe ou faz (competências) é uma representação estática, indireta e generalista das competências humanas e não se baseia nas práticas de trabalho. Daley (2002) diz que as competências requeridas para a execução de um trabalho não permanecem fixas e, nesse caso, é importante que os mapeamentos ou modelos de competência, frutos de um processo de identificação de competências, sejam revisitados constantemente. O problema reside no fato de que os custos associados a tais esforços, tanto em termos de tempo quanto de dinheiro, fazem com que essa prática seja negligenciada.

Lindgren, Stenmark e Ljungberg (2003) destacam que competências estão associadas a processos de mudança e resolução de problemas, um contexto típico de organizações baseadas em conhecimento. Portanto devem ser vistas de forma dinâmica, emergente e envolvidas com a prática. Isso leva a crer que um processo de identificação de tais competências deve observar essas características.

Acredita-se que outro aspecto a ser considerado é o caráter latente da explicitação das competências presente no contexto de aplicação da técnica de entrevista. Latente porque o colaborador pode não se sentir à vontade de explicitá-

las ou até mesmo não ter a total consciência das competências requeridas para a execução das atividades. Essa situação tende a produzir um mapeamento de competências que pode não condizer com a realidade atual da organização.

Nesse sentido, novas formas de identificação ou descoberta de elementos constituintes do capital humano da organização se fazem necessárias.

4.1.3 FORMAS DE IDENTIFICAÇÃO DE CAPITAL HUMANO

Para Becker (1964) o capital humano não é apenas o resultado da educação formal, mas inclui experiências e práticas de aprendizagem que ocorrem no ambiente de trabalho. Portanto, atividades executadas pelos colaboradores dentro da organização e os insumos textuais eletrônicos decorrentes do uso de diversas ferramentas de tecnologia de informação e comunicação podem ser possíveis fontes de identificação de elementos de capital humano.

Algumas atividades típicas de colaboradores de organizações de conhecimento que geram insumos passíveis de verificação e identificação de tais elementos são:

- Pesquisas realizadas em *search engines* como Google, Yahoo, etc.: essa é uma atividade típica de colaboradores que trabalham em organizações baseadas em conhecimento, já que *search engines* se tornaram a principal fonte de pesquisa para a resolução de problemas. Nesse tipo de atividade, conhecimentos e interesses relacionados ou não às tarefas que o colaborador executa são explicitados através da digitação dos termos de busca, os quais podem ser capturados e analisados posteriormente.

- Envio e recebimento de *e-mails*: essas são atividades típicas de qualquer tipo de organização. *E-mails* são poderosas ferramentas de comunicação dentro da organização. Permitem a troca de informações que vão desde avisos de cunho administrativo até informações sobre projetos, formas de resolução de problemas e interesses em comum. Nesse tipo de atividade, conhecimentos relacionados ou não às tarefas que o colaborador executa são explicitados e também podem ser capturados e analisados posteriormente.

- Utilização de ambientes de colaboração como Wiki: essa também é uma atividade que lentamente está sendo introduzida em organizações baseadas em conhecimento como uma forma de produzir e compartilhar conhecimento coletivamente. Explicita fortemente conhecimentos muito relacionados às práticas de trabalho na organização. O conteúdo inserido nesse tipo de ambiente também pode ser analisado.

- Uso de blog pessoal: o blog difere de um ambiente de colaboração como o Wiki porque permite que conteúdo seja inserido de forma individual e também com cunho pessoal e informal. Nesse sentido, se caracteriza como uma fonte de análise de conhecimentos do colaborador.

Todas as atividades citadas acima geram insumos textuais relacionados às práticas de trabalho e são, como discutido anteriormente, possíveis fontes de identificação de elementos constituintes do capital humano.

4.1.4 MODELO DE DESCOBERTA DE CONHECIMENTOS E INTERESSES

A utilização de mecanismos que auxiliem no processo de identificação ou mapeamento do capital humano na organização é uma necessidade. O modelo proposto nesse artigo vai ao encontro dessa necessidade, pois visa a descobrir conhecimentos e interesses baseando-se em insumos textuais eletrônicos produzidos a partir das práticas de trabalho dos colaboradores numa abordagem semiautomática.

A escolha dos elementos conhecimento e interesse se deu em função de um teor menor de subjetividade em relação aos outros elementos constituintes do conceito de capital humano, como habilidades e atitudes. Acredita-se que os elementos conhecimento e interesse são possíveis de serem identificados através do uso, no dia a dia de trabalho, de algumas ferramentas como e-mail, ferramentas de busca, etc. Essa afirmação é fundamentada no trabalho de Stenmark (2000) que afirma que a exploração das ações cotidianas de um usuário na intranet em um dia de trabalho, de uma forma não intrusiva, pode conduzir ao descobrimento de padrões invisíveis.

É comum que essas ferramentas gerem uma saída textual, aqui denominada insumo textual, que é armazenada e é, conseqüentemente, passível de ser

analisada. Segundo Konchady (2006) mais de 80% das informações de uma organização estão armazenadas num formato textual não estruturado. Esta informação geralmente está distribuída entre muitos computadores dentro da organização e, diferentemente dos dados estruturados, não possui um repositório único para armazená-la.

O modelo de descoberta proposto neste artigo se dá através de uma abordagem tecnológica apoiada num ferramental da Engenharia do Conhecimento, que utiliza um processo de KDT (*Knowledge Discovery from Text*) cujas etapas são sugeridas por Feldman e Sanger (2007) e são as seguintes: coleta de dados, extração de informação, mineração e análise da informação. A etapa de mineração é apoiada por um conhecimento de domínio (*background knowledge*) representado formalmente via uma ontologia de tópicos ou áreas de conhecimento. A descoberta de conhecimentos propriamente dita se dá através de um algoritmo de *matching* e inferência nessa ontologia. O algoritmo usa como entrada os insumos textuais coletados oriundos das práticas de trabalhos dos colaboradores.

O resultado da aplicação do modelo é uma lista de tópicos ou termos que representam conhecimentos ou interesses do colaborador e podem servir de insumo para o processo de identificação do capital humano da organização. A periodicidade de aplicação do modelo pode ser variada, o que permitirá a descoberta de conhecimentos e interesses atualizados e possivelmente pertinentes às práticas de trabalho na organização. A Figura 01 apresenta o fluxo de informação do modelo proposto.

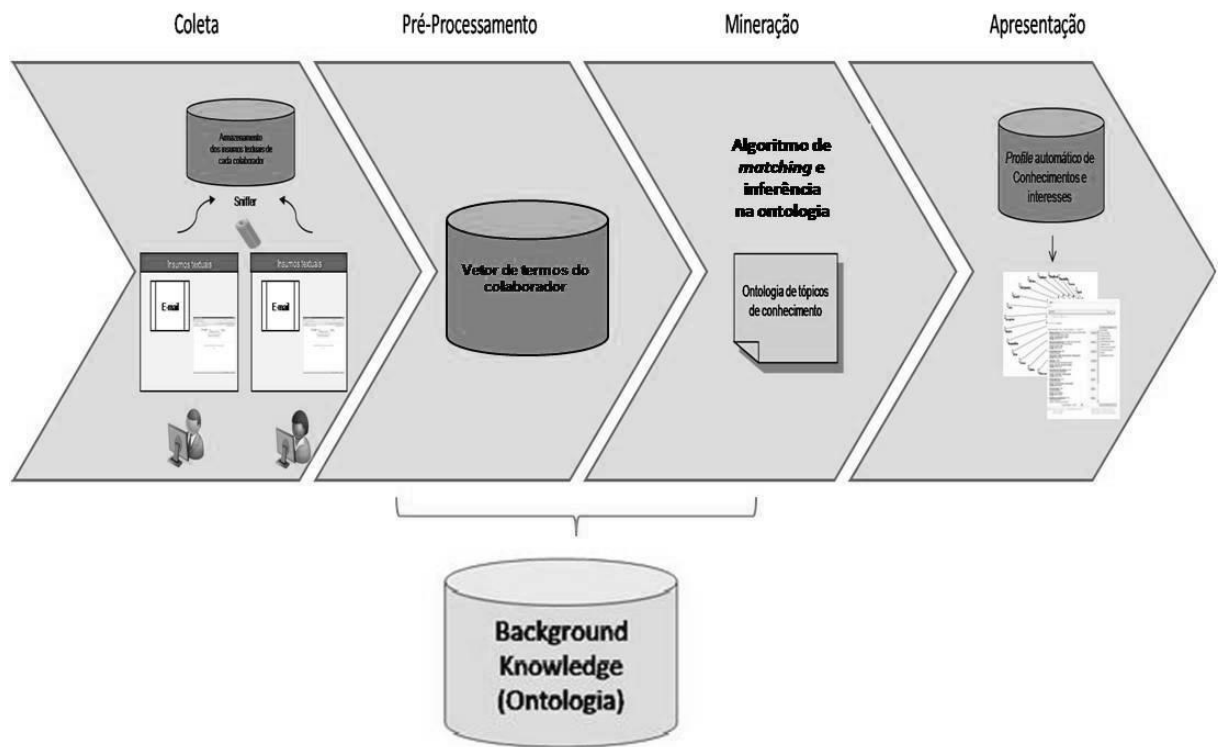


Figura 1 - Fluxo de informação do modelo proposto
Fonte: Elaboração dos autores

Na etapa de coleta, os insumos textuais produzidos e armazenados no computador pessoal de cada colaborador durante seu trabalho, tais como e-mails e termos digitados em ferramentas de busca, são coletados automaticamente. Optou-se por essas duas fontes de dados porque se acredita que sejam fontes que expressem conhecimentos e interesses mais subjetivos.

Na etapa de pré-processamento os insumos são preparados para a criação do vetor dos termos mais frequentes presentes na coleção de documentos do colaborador. Aqui a ontologia de tópicos de conhecimento é utilizada no processo de identificação dos termos ("tokenização") para dar consistência aos termos (*tokens*). O vetor de termos é utilizado como insumo para o algoritmo de descoberta ou mineração de conhecimentos e interesses, utilizado na etapa de mineração.

A etapa a mineração é feita única e exclusivamente através da ontologia de tópicos de conhecimento. Um algoritmo de análise e inferência nessa ontologia utiliza os termos do vetor do colaborador e determina a descoberta de conhecimentos e interesses.

Na etapa de apresentação, os conhecimentos e interesses descobertos são apresentados na forma de uma lista ou *profile* e servem como insumo para a tomada de decisão na organização.

A seguir serão detalhados alguns itens importantes do modelo proposto:

a) Vetor de termos

Para cada colaborador são criados três vetores com os termos mais frequentes e a respectiva frequência de ocorrência na sua coleção de documentos (emails enviados e recebidos e termos de busca digitados). Um vetor contém os termos mais frequentes encontrados na coleção de e-mails enviados e recebidos. O outro vetor contém os termos mais frequentes encontrados nos termos de busca digitados e o último vetor contém os termos mais frequentes encontrados na coleção de e-mails e termos de busca digitados conjuntamente.

Os três vetores criados a partir de coleções textuais diferentes (*e-mails*, termos de busca digitados e ambos) são necessários, uma vez que o modelo proposto visa a descobrir dois elementos distintos - conhecimentos e interesses - e não é possível determinar de antemão qual coleção de documentos representa melhor os conhecimentos existentes e qual coleção representa interesse. Com a existência de três vetores para cada colaborador, é possível aplicar o algoritmo de *matching* e inferência na ontologia, utilizando os termos de cada vetor de maneira independente e assim descobrir também qual coleção de dados representa com mais acurácia os conhecimentos e interesses de cada colaborador.

Os termos mais frequentes de cada vetor são determinados da seguinte forma:

Passo 1 - Ontologia para coleção: cada termo de uma instância da ontologia é procurado na coleção de documentos específica para criação do vetor e contabilizado. Termos com baixa frequência são eliminados.

Passo 2 - Coleção para a ontologia: termos encontrados na coleção de documentos do colaborador que não foram detectados no passo 1, mas possuem uma frequência alta na coleção, também são considerados e fazem parte do vetor.

b) Ontologia do modelo

A ontologia utilizada nesse modelo tem um papel vital. É nos termos presentes nas instâncias das classes que o algoritmo de descoberta ou mineração navegará,

encontrando os termos desejados, seus relacionamentos, e descobrindo, assim, novos conhecimentos e interesses. Essa seção descreve a ontologia, sua metodologia de construção, o processo de inserção das instâncias e estabelecimento dos relacionamentos e seus respectivos pesos e, por último, a atualização da ontologia.

A ontologia utilizada nesse trabalho foi concebida segundo os passos sugeridos por Noy e McGuinness (2001). Ela é denominada ontologia de tópicos de conhecimento porque representa o conhecimento de domínio de áreas de conhecimento e suas relações. Para isso, foi criada uma estrutura com quatro classes (Área, Subárea, Assunto e Termos relacionados) que representam conceitos relacionados a áreas ou tópicos de conhecimento.

As classes estão relacionadas através de um relacionamento “parte-de”, em que a primeira classe representa o conceito de área do conhecimento geral, a segunda classe representa o conceito de subárea do conhecimento, a terceira classe representa o conceito de assuntos abordados em uma subárea do conhecimento e a quarta classe representa os termos relacionados aos assuntos e que não podem ser classificados como área, subárea ou assunto.

As instâncias da ontologia são inseridas manualmente pela organização, ou seja, a organização deve escolher as áreas de conhecimento que deseja descobrir conhecimentos e interesses de colaboradores e estabelecer os relacionamentos existentes entre uma determinada área e subáreas, uma determinada subárea e assuntos e um determinado assunto e termos relacionados.

As instâncias e relacionamentos dos três primeiros níveis de conceitos da ontologia podem ser determinados pelo que é preconizado na literatura da área de conhecimento em questão. O exemplo abaixo mostra as subáreas da área Ciência da Computação e os assuntos da subárea Inteligência Artificial existentes na literatura dessa área.

Área: Ciência da Computação

Subárea: Inteligência Artificial, Engenharia de Software, Banco de Dados, etc.

Assunto: Recuperação de informação, Extração de informação, Robótica, etc.

As instâncias do quarto nível (Termo relacionado) da ontologia e seus relacionamentos com as instâncias do terceiro nível da ontologia (Assunto) também podem ser determinados em parte pelo que é encontrado na literatura, mas essa

relação não é tão explícita e de senso comum como as instâncias e relações dos níveis anteriores.

Por isso, um algoritmo para cálculo de coocorrência entre as entidades (termos, *tokens*) do corpus ou coleção de documentos textuais (insumos textuais produzidos pelos colaboradores durante seu trabalho) que está sendo analisado é aplicado, visando ajudar a descoberta de relações latentes ou não explícitas entre um assunto e os termos mais relacionados a esse assunto, auxiliando consequentemente a inserção dos termos e relacionamentos com determinado assunto na ontologia. Alguns dos algoritmos de coocorrência disponíveis para uso são: teste t, Chi-square (χ^2) e Z score (Manning & Schütze, 1999).

Além da inserção das instâncias e relacionamentos com outras instâncias, o peso ou medida que descreve a força ou grau do relacionamento de uma instância com outra instância é estabelecido. Esse peso pode variar de 0 a 1.

Para os primeiros dois níveis de relacionamento da ontologia (Área e Subárea, Subárea e Assunto) é estabelecido que cada relacionamento tenha o mesmo peso, ou seja, divide-se o peso máximo (1) da instância pelo seu número de relacionamentos. A princípio, acredita-se que uma área é composta de todas as suas subáreas, não existindo uma subárea que se sobressaia, ou seja, que tenha uma força de relacionamento com a sua área maior que as demais subáreas. A mesma justificativa vale para uma subárea e seus assuntos. Porém essa é uma decisão estratégica da organização, ou seja, se uma determinada área ou subárea tem mais importância para o negócio, os pesos podem ser calibrados.

O cálculo de coocorrência entre entidades, citado anteriormente como um mecanismo de descoberta de relações não explícitas entre entidades, e executado previamente à criação das instâncias da ontologia, objetiva identificar entidades que coocorrem num mesmo *corpus* e também a força dessa coocorrência, que é uma medida que determina o quanto duas entidades estão correlacionadas, também chamada de correlação. Essa medida é utilizada para atribuição dos pesos proporcionais das relações entre as instâncias do terceiro e quarto nível da ontologia, respectivamente Assunto e Termo relacionado.

A Figura 2 demonstra uma atribuição hipotética de pesos numa amostra de instâncias relacionadas de uma ontologia de tópicos de conhecimento da área da Ciência da Computação.

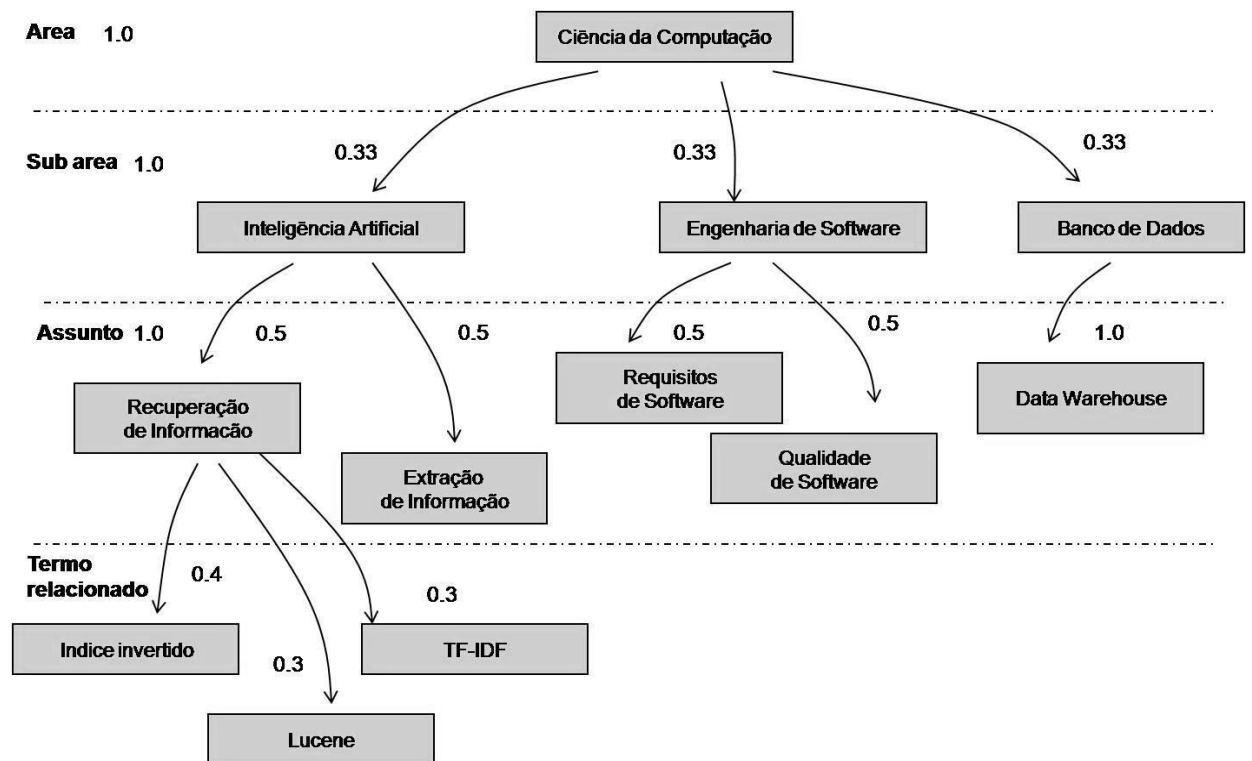


Figura 2 - Amostra de instâncias relacionadas e seus respectivos pesos
 Fonte: Elaboração dos autores

A maior parte do processo de atualização da ontologia é manual, ou seja, as instâncias de Área, Subárea, Assunto e Termos relacionados e seus relacionamentos são definidos e inseridos pela organização. Termos com alta frequência presentes no vetor do colaborador que não estão presentes na ontologia devem servir de insumo para a atualização da mesma. Uma periodicidade de atualização é estabelecida. A cada atualização os tópicos de conhecimento (área, subárea, assunto e termos relacionados) e seus relacionamentos são revistos.

c) Algoritmo de matching e inferência na ontologia

O algoritmo tem o objetivo de descobrir novos conhecimentos ou interesses a partir dos termos do vetor do colaborador. Isso não significa que cada termo do vetor e sua respectiva frequência sejam analisados isoladamente e com isso considere-se que exista um conhecimento ou interesse no dado termo. Ao contrário disso, o algoritmo tem a função de descobrir para cada termo encontrado no vetor do

colaborador as relações com outros termos e seus respectivos pesos presentes na ontologia e assim inferir que um determinado conhecimento ou interesse existe.

Cada termo do vetor do colaborador é pesquisado em todas as instâncias da ontologia de tópicos de conhecimento. O termo pode ser encontrado em uma instância de qualquer conceito ou classe da ontologia (Termo relacionado, Assunto, Subárea ou Área) ou pode não ser encontrado. Se o termo for encontrado, ele vai estar relacionado com no mínimo um termo, já que todas as instâncias da ontologia possuem no mínimo um relacionamento.

A partir do momento em que o termo é encontrado na ontologia, a abordagem de análise ou inferência para a descoberta de conhecimento ou interesse a partir do termo encontrado pode acontecer de duas maneiras: *bottom-up* ou *middle-down-up*.

A justificativa para esses dois tipos de abordagem de análise ou inferência reside no pressuposto de que um termo existente no vetor do colaborador, no qual somente termos com alta frequência estão presentes, ao ser encontrado em alguma instância da ontologia não é por si só um indicativo de conhecimento ou interesse do colaborador naquele termo ou tópico de conhecimento, seja ele encontrado numa instância da classe Termo Relacionado, Assunto, Subárea ou Área. É necessária a análise do relacionamento do termo encontrado em uma determinada instância com outros termos (instâncias) relacionados abaixo ou acima da classe onde o termo foi encontrado.

Na abordagem *bottom-up* o termo do vetor do colaborador é encontrado no último nível da ontologia, a classe Termo Relacionado. Nessa abordagem a inferência é feita considerando-se a relação do conceito filho para o conceito pai, ou seja, da instância de nível hierárquico inferior (*bottom*) para o superior (*up*).

O processo de inferência acontece em três etapas: a primeira etapa analisa o relacionamento entre as instâncias da classe Termo Relacionado e Assunto; a segunda etapa analisa o relacionamento entre as instâncias da classe Assunto e Subárea; a terceira etapa analisa as instâncias da classe Subárea e Área.

Na primeira etapa, a frequência de cada termo do vetor do colaborador encontrado entre as instâncias da classe Termo Relacionado que se relaciona com uma determinada instância da classe Assunto, que é a classe imediatamente superior na hierarquia, é multiplicada pelo peso da relação onde o termo foi

encontrado. Uma instância da classe Termo relacionado pode estar relacionada com mais de uma instância da classe Assunto e cada relação pode ter um peso diferente.

O somatório do produto da frequência de todos os termos do vetor do colaborador encontrados nas instâncias da classe Termo relacionado (último nível da ontologia) que se relacionam com uma mesma instância da classe Assunto (nível imediatamente superior) pelo peso dessa relação representa, nesse artigo, o grau de conhecimento ou interesse naquele tópico de conhecimento, nesse caso, um Assunto. A seguinte equação determina essa medida:

$$\sum t_f \cdot r_w$$

onde t_f é a frequência do termo no vetor do colaborador e r_w é o peso da relação desse termo com outro termo na ontologia.

A segunda etapa analisa o relacionamento entre a instância da classe Assunto, que obteve uma medida ou grau de conhecimento/interesse em função do processo realizado na primeira etapa, com as instâncias da classe imediatamente superior na hierarquia, a classe Subárea. Nessa etapa, cada instância da classe Assunto que possui uma medida ou grau de conhecimento/interesse calculado na primeira etapa tem esse valor multiplicado pelo peso das relações que possui.

O somatório do produto do grau ou medida (peso) de conhecimento/interesse de todos os assuntos que se relacionam com uma determinada Subárea pelo peso dessa relação representa o grau de conhecimento ou interesse naquele tópico de conhecimento, nesse caso, uma Subárea. A seguinte equação determina essa medida:

$$\sum c_i w \cdot r_w$$

Na terceira e última etapa é analisado o relacionamento entre a instância da classe Subárea, que obteve uma medida ou grau de conhecimento/interesse em função do processo realizado na segunda etapa, com a única instância da classe Área a que a mesma está relacionada. Uma instância da classe Subárea só pode estar relacionada com uma instância da classe Área.

O somatório do produto do grau ou medida (peso) de conhecimento/interesse de todas as subáreas que se relacionam com uma determinada Área pelo peso dessa relação representa nesse artigo o grau de conhecimento ou interesse naquele tópico de conhecimento, nesse caso, uma Área.

A abordagem *middle-down-up* acontece quando o termo encontrado na ontologia não pertence ao último nível da ontologia, a classe Termo relacionado. Nesse caso o termo pode ser encontrado em instâncias das classes Assunto, Subárea ou Área. Nessa abordagem a inferência é feita considerando-se as relações abaixo (*down*) e acima (*up*) a partir da classe onde o termo foi encontrado.

Ao encontrar um termo numa instância de classe da ontologia, que não a classe Termo relacionado, é necessário avaliar primeiramente todos os relacionamentos dessa instância com as instâncias de classes abaixo da classe onde o termo foi encontrado (*middle-down*) e posteriormente com as instâncias de classes acima (*middle-up*).

A análise *middle-down* também acontece em etapas, sendo que o número das etapas depende da classe onde o termo for encontrado. Se o termo for encontrado numa instância da classe Assunto, uma etapa de análise é necessária, pois somente os relacionamentos da instância de Assunto com as instâncias de Termo relacionado são analisados. Se o termo for encontrado numa instância da classe Subárea, duas etapas de análise são necessárias, pois os relacionamentos da instância de Subárea com as instâncias de Assunto e posteriormente de Assunto com Termo relacionado são analisados. A mesma lógica de análise por etapas ocorre se o termo for encontrado em uma instância da classe Área.

Essa análise se justifica porque não é possível assumir isoladamente que um termo do vetor do colaborador, quando encontrado em alguma instância das classes Assunto, Subárea ou Área, seja indicativo, em algum grau, de conhecimento ou interesse. É necessária a verificação da existência dos termos presentes nos relacionamentos abaixo de onde o termo foi encontrado no vetor do colaborador. A presença desses termos no vetor do colaborador e o peso das suas relações com o termo analisado contribuem para que o mesmo seja considerado, em algum grau, um conhecimento ou interesse.

Em todas as etapas o processo de análise é o mesmo. São analisados todos os relacionamentos da instância onde o termo foi encontrado com outras instâncias da classe imediatamente inferior. Os termos das instâncias relacionadas são pesquisados no vetor do colaborador, caracterizando um processo inverso ao apresentado nas abordagens, ou seja, a pesquisa (*matching*) parte dos termos das instâncias relacionadas à instância onde o termo foi encontrado para o vetor do

colaborador. Para cada termo encontrado no vetor do colaborador aplica-se o mesmo processo explicado na abordagem *bottom-up*: multiplica-se a frequência do termo pelo peso da relação.

Nessa análise, o grau ou medida de conhecimento ou interesse do termo analisado é dado pelo somatório da frequência do termo no vetor do colaborador (t_f) multiplicado pelo peso da relação desse termo (r_w) com o termo analisado na ontologia acrescido de um fator “f” relacionado à frequência do próprio termo analisado encontrado na ontologia.

O fator “f” tem o objetivo de incrementar o grau ou medida de conhecimento ou interesse que o termo que está sendo analisado possui. Assume-se que, se o termo analisado está presente no vetor do colaborador e alguns de seus termos filhos (fruto da relação com as instâncias da classe imediatamente inferior) também estão, a frequência do próprio termo analisado deve ser considerada na equação. A seguinte equação representa essa idéia:

$$\sum t_f \cdot r_w + f$$

Terminada a análise *middle-down* do termo analisado, entra em execução a análise *middle-up*, cujo processo é exatamente igual ao descrito na segunda e terceira etapa da abordagem *bottom-up*, ou seja, são analisados os relacionamentos entre a instância do termo em análise, que já obteve uma medida ou grau de conhecimento/interesse em função da análise *middle-down*, com as instâncias da classe imediatamente superior na hierarquia e assim sucessivamente.

4.1 Delimitações do modelo

Os insumos textuais analisados no modelo serão: e-mails e termos digitados em ferramentas de busca como *Google*, *Yahoo*, *Live Search*, etc. O conjunto de insumos textuais relacionado a cada colaborador é identificado prévia e automaticamente no momento da coleta, ou seja, não existe um algoritmo para detecção da autoria do insumo textual a partir de um único e grande repositório. A ontologia é atualizada pela organização, porém com subsídio de informação de um processo semiautomático que indica possíveis termos e relacionamentos.

4.1.5 Cenário de aplicação do modelo

Essa seção descreve o cenário ideal e o fluxo de aplicação do modelo. O cenário deve ser característico de uma organização baseada em conhecimento, onde os colaboradores possuam alto nível de formação educacional e experiência e exista uma cultura forte de compartilhamento de conhecimento.

A amostra utilizada para a aplicação do modelo deve contar com grupo de aproximadamente 10 colaboradores. Os colaboradores escolhidos para participarem da aplicação devem estar diretamente relacionados à área de conhecimento escolhida para a descoberta de conhecimentos e interesses e produzirem insumos textuais no local de trabalho em quantidade relevante para a aplicação do modelo. Recomenda-se que o período de coleta de dados seja de seis meses a um ano. O fluxo de aplicação pode ser observado na Figura 03.

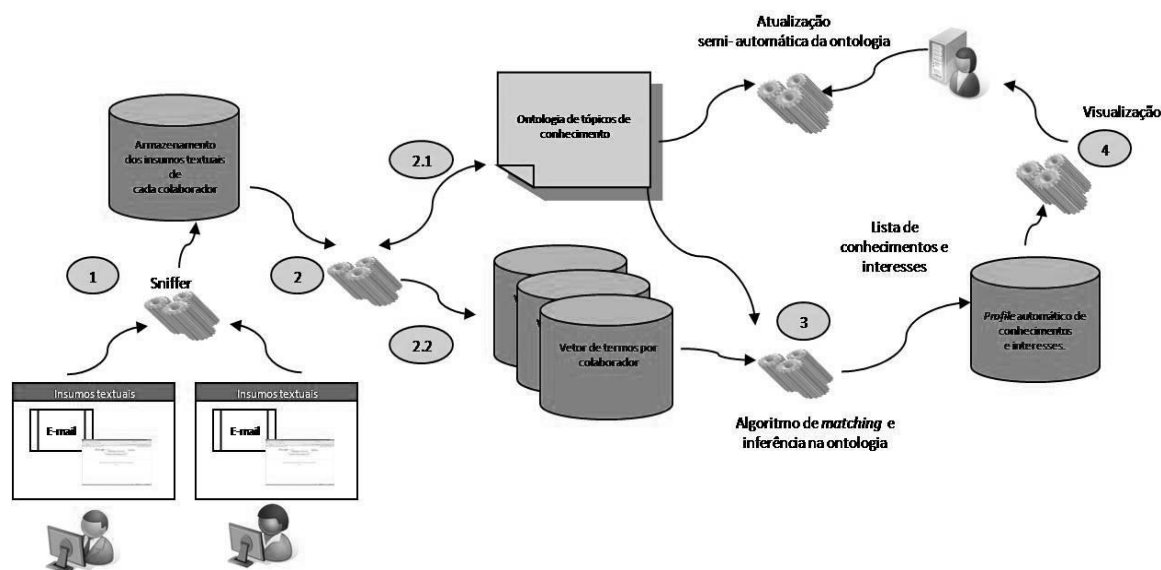


Figura 3 - Fluxo da aplicação do modelo

Fonte: Elaboração dos autores

O fluxo 1 da Figura 3 demonstra a etapa de coleta dos dados. Os e-mails enviados e recebidos e os termos digitados em ferramentas de busca pelo grupo de colaboradores escolhidos para participarem da aplicação do modelo são coletados. Em cada estação de trabalho do colaborador, uma aplicação do tipo *sniffer* (aplicação capaz de interceptar e registrar o tráfego de dados em uma rede de

computadores) intercepta pacotes *SMTP* e *HTTP* das *url* dos sites de buscas previamente configurados e armazena os dados num repositório onde o dado e a sua origem (colaborador) são identificados.

Após ser coletado e armazenado, o *corpus* (coleção de insumos textuais) de cada colaborador é pré-processado para a execução de duas atividades, onde uma antecede a outra, o que pode ser visto no fluxo 2.1 e 2.2 da Figura 3. O fluxo 2.1 mostra o processo de cálculo de coocorrência que gera os termos relacionados que podem ser utilizados para atualização do último nível da ontologia. O fluxo 2.2 mostra a criação dos vetores de termos frequentes do colaborador. Esse processo é aplicado para o conjunto de insumos textuais do tipo e-mail, termos de busca e ambos os tipos conjuntamente, gerando os três respectivos vetores. O processo para criação dos vetores ocorre em dois passos: no primeiro passo cada termo presente nas instâncias da ontologia é pesquisado no conjunto dos insumos textuais e contabiliza-se a frequência com que o termo for encontrado; no segundo passo é feita a “tokenização”, a remoção das *stop words* e cada termo é contabilizado. Somente termos com frequência significativa vão para o vetor de termos do colaborador. Em ambos os casos, a frequência do termo é relativa em relação à frequência desse termo no conjunto geral dos insumos textuais de todos os colaboradores. O segundo passo foi proposto porque se acredita que muitos termos não presentes na ontologia serão encontrados e eles podem ser, por si só, dependendo da sua frequência, um indicativo de conhecimento ou interesse naquele termo. Além disso, termos com alta frequência que não estão presentes na ontologia devem ser utilizados pela organização no processo de atualização da ontologia. A periodicidade de geração dos vetores é determinada pela organização. A cada nova geração, considera-se todo o conjunto de insumos textuais do colaborador.

O fluxo 3 mostra a execução do algoritmo de *matching* e inferência na ontologia. Ele tem como entrada o vetor de termos do colaborador e a ontologia de tópicos de conhecimento, pesquisando-se cada termo do vetor na ontologia (*matching*); a partir disso infere-se qual o grau ou medida de conhecimento ou interesse em outros termos relacionados com o termo em questão. Cabe ressaltar que o mesmo algoritmo é aplicado de maneira independente para cada um dos vetores do colaborador. A saída do algoritmo é uma lista de termos ou tópicos em que o colaborador possui conhecimento, ou interesse, ou ambos. Cada tópico da lista estará associado a um grau, medida ou percentual que representa o nível de

conhecimento/interesse no tópico. A lista de tópicos caracteriza o perfil (*profile*) de conhecimentos ou interesses do colaborador na organização. No fluxo 4 os resultados podem ser visualizados pela organização.

4.1.6 Resultados esperados

Espera-se que a aplicação do modelo proposto produza uma lista de termos ou tópicos de conhecimento e interesses que reflita os conhecimentos e interesses reais dos colaboradores. Espera-se, além disso, a verificação de qual tipo de insumo textual melhor reflete os conhecimentos desses dos colaboradores.

4.1.7 5 CONSIDERAÇÕES FINAIS

Nesse artigo, apresentou-se um modelo para descoberta semiautomática de conhecimentos e interesses. O modelo objetiva contribuir ou facilitar o processo de identificação do capital humano, fornecendo uma lista de tópicos ou termos que representam os conhecimentos e interesses dos colaboradores de um organização, a partir dos insumos textuais produzidos pelos próprios colaboradores. Com o mapeamento do capital humano bem definido e atualizado, várias ações de gestão podem ser tomadas, tais como a alocação eficiente dos colaboradores dentro da organização, acompanhamento da progressão profissional dos colaboradores, treinamento para aquisição dos conhecimentos exigidos pela organização, etc.

REFERÊNCIAS

- Becker, G. S. (1964). *Human capital: a theoretical and empirical analysis, with special reference to education*. Chicago: The University of Chicago Press.
- Daley, D. M. (2002). *Strategic human resource management: People and performance management in the public sector*. Upper Saddle River, NJ: Prentice Hall.
- Drucker, P. (1999, March-April). Managing Oneself. *Harvard Business Review*, 77(2):64-74.
- Edvinsson, L., & Malone, M. S. (1998). *Capital intelectual*. São Paulo: Makron Books.

- Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.
- Konchady, M. (2006). *Text Mining Application Programming*. Boston: Charles River Media.
- Lindgren, R., Stenmark, D., & Ljungberg, J. (2003). Rethinking Competence Systems for Knowledge-Based Organizations. *European Journal of Information Systems*, 12(1), 18–29.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.
- Nonaka ,I., & Takeuchi, H. (1997). *Criação do Conhecimento na Empresa*. Rio de Janeiro: Campus.
- Noy, F. N., & Guinness, D. L. (2001). *Ontology development 101: a guide to create your first ontology*. Recuperado em 10 outubro, 2013, de <http://ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.doc>.
- Sandberg, J. (2000). Understand Human Competence at Work: An Interpretative Approach. *Academy of Management Journal*, 43(1), 9-25.
- Sveiby, K-E. (1999). *Welcome to the knowledge organization*. Recuperado em 10 outubro, 2013, de <http://www.sveiby.com/articles/K-era.htm>
- Stenmark, D. (2000, January). Turning Tacit Knowledge Tangible. *Proceedings of the Hawaii International Conference on System Sciences (HICSS33)*, Wailea Maui, Hawaii, 33.
- Stewart, T. A. (2002). *A riqueza do conhecimento: O capital intelectual e a nova organização*. Rio de Janeiro: Campus.