A recuperação semântica da informação no contexto do controle externo

Márcia Martins de Araújo Altounian
é servidora do Tribunal de Contas da União, bacharel em Biblioteconomia e Documentação pela UnB, com especialização em Gestão do Conhecimento pelo Instituto Blaise Pascal e Arquitetura e Organização da Informação pela UFMG.

Beatriz Pinheiro de Melo Gomes
é servidora do Tribunal de Contas da União, bacharel em Biblioteconomia e Documentação pela UnB, com especialização em Gestão Estratégica do Conhecimento e Inteligência Empresarial pela PUC/Paraná.

RESUMO

Nos sistemas de recuperação da informação (SRI), as técnicas de base sintática têm sido suplantadas pela crescente exploração das técnicas de recuperação semântica, que possibilitam a compreensão dos conceitos em seu contexto e finalidade. Algumas tecnologias têm contribuído para essa realidade, como a marcação semântica dos dados, utilizada na web semântica, o processamento de linguagem natural e as redes neurais. O tesauro também apresenta-se como um componente semântico que impacta no desempenho dos SRI. Tesauros são ferramentas da linguagem artificial em um domínio específico, formados por um sistema de conceitos relacionados entre si. Este artigo apresenta a aplicação do tesauro do TCU, denominado Vocabulário de Controle Externo, em alguns sistemas de informação corporativos.

Palavras-chave: Semântica. Recuperação da informação. Tesauro.

1. INTRODUÇÃO

A recuperação da informação (RI) é um campo de interesse comum da ciência da computação e da ciência da informação que se preocupa em desenvolver e estudar os aspectos relativos a eficiência e eficácia das buscas em um sistema de informação, de modo que os resultados sejam

coerentes com sua expressão de busca e, sobretudo, relevantes para o usuário do sistema.

Efetuar buscas em grandes repositórios de informações não estruturadas e não padronizadas torna-se uma tarefa árdua, levando muitas vezes a ambiguidades, a conteúdos fora de contexto e até mesmo a não recuperação da informação desejada. As ferramentas de busca, ao fazerem uso da linguagem natural, necessitam de conhecimento sobre o significado das expressões utilizadas e das relações entre elas, o que possibilita contextualização e tratamento de fenômenos linguísticos que afetam a qualidade da recuperação.

Para o processo de classificação e recuperação de informação, vários métodos automáticos vêm sendo desenvolvidos com o objetivo de obter respostas relevantes para qualquer pesquisa realizada. Até o momento, a maioria deles baseia-se em aspectos sintáticos e estatísticos, levando em consideração frequência e distribuição de palavras presentes em documentos. Por se basear nos aspectos sintáticos da informação, a classificação realizada por esses métodos ainda está muito distante, em termos de qualidade, da realizada com base na indexação dos assuntos dos documentos feita por especialistas.

Recentemente, vêm sendo explorados novos métodos de classificação automática de documentos baseados em aspectos semânticos da linguagem que se referem ao significado dos termos em seu contexto e finalidade. Uma dessas iniciativas consiste no uso de instrumentos de representação de relacionamentos semânticos e conceituais como os tesauros, que se caracterizam por representar domínio do conhecimento e objetivam solucionar problemas de ambiguidade inerentes às palavras usadas cotidianamente.

O Vocabulário de Controle Externo, tesauro do TCU, foi desenvolvido com o objetivo de padronizar o tratamento de conteúdos especializados e contribuir com a recuperação semântica no contexto do controle externo, além de conferir maior agilidade e precisão às pesquisas nos sistemas de informação do Tribunal.

2. SEMÂNTICA

O termo "semântica", tradicionalmente estudado na linguística e na filosofia, vem sendo amplamente usado nas mais variadas áreas de atuação, especialmente nas relacionadas à tecnologia da informação. A chamada web semântica (WS), que tem origem na expansão da web e nas limitações dos instrumentos de busca baseados em sintaxe, é certamente um dos motivos para essa associação.

Mas o que é semântica? A palavra tem origem no termo grego *semantiké*, e muitas acepções podem ser encontradas de acordo com a perspectiva utilizada, pois há, entre outras vertentes, semântica textual, cognitiva, lexical, formal e argumentativa. Em geral, todas convergem num mesmo ponto: estudam significado, ou significação. A ampla variedade de possibilidades atesta que o estudo do significado pode ser feito de vários ângulos.

Três propriedades básicas destacam-se na semântica – são elas a sinonímia, a antonímia e a polissemia. A sinonímia é a divisão da semântica que estuda a relação entre expressões linguísticas que têm o mesmo sentido. Por

exemplo, "garota" e "menina" são palavras substantivas que possuem um mesmo significado, remetendo à figura de uma jovem; assim também acontece com verbos como "renunciar", "recusar" e "rejeitar", que transmitem a ideia de repulsa. Ressaltando-se a rara ocorrência de sinonímia perfeita, pode-se afirmar que sinonímia é a relação entre palavras e expressões que possuem sentido e significado comuns.

Se por um lado a sinonímia estuda as palavras com significados semelhantes na língua, a antonímia trata do estudo das palavras que indicam sentidos opostos. Na mesma linha do exemplo anterior, podemos citar os substantivos "garota" e "senhora", bem como os verbos "renunciar" e "aceitar" – tratam-se todos de termos com significação oposta.

Também é possível que uma mesma palavra assuma significados diferentes –nesse caso, o contexto em que estiver inserida ditará o seu sentido. Um bom exemplo do cotidiano é a palavra "manga", da qual rapidamente podemos pensar em duas acepções bastante diferentes: uma fruta e uma parte de peça do vestuário. Esse é um exemplo comum de polissemia, termo que, formado pelo prefixo "poli" ("muitos") e pelo sufixo "semos" ("significados"), é a parte da semântica que estuda as significações que uma palavra assume em determinado contexto linguístico.

Há ainda outras duas propriedades semânticas, na mesma linha de estudo da significação, que merecem ser citadas: a homonímia e a paronímia. A homonímia estuda a relação de duas ou mais palavras que, apesar de terem significados diferentes, têm a mesma forma e o mesmo som – é o caso de termos como concerto/conserto e rio/rio (substantivo/verbo), entre tantos outros. A paronímia estuda as particularidades de palavras que, ao contrário, têm grafia e pronúncia semelhantes, mas significados diferentes – são exemplos: eminente/iminente e absolver/absorver.

Finalmente, a semântica estuda ainda as propriedades de denotação e conotação das palavras. Denotação é a propriedade que uma palavra tem de limitar-se ao seu próprio conceito – por exemplo, o termo "estrelas" em "as estrelas do céu". Conotação é a propriedade que uma palavra tem de ampliar-se no seu campo semântico, dentro de um contexto, gerando várias possibilidades interpretativas – por exemplo, o mesmo termo "estrelas" em "as estrelas do cinema".

Quando transportamos os conceitos aqui explorados para o que hoje conhecemos como WS, fica fácil perceber que estamos falando de promover melhorias nos processos de representação e recuperação da informação na web. Desde 1990, a web é caracterizada por usar linguagens de marcação que objetivam a apresentação e a leitura por pessoas e por mecanismos de busca baseados em algoritmos com orientação à sintaxe. O uso da semântica pode ampliar a possibilidade de associações dos documentos a seus significados por meio de metadados descritivos. Portanto, a questão do significado na semântica é fundamental à WS.

Das diversas linhas de estudo relacionadas à semântica, a da semântica formal parece ser a mais pertinente à tecnologia da informação. Três aspectos principais permeiam os estudos sobre semântica formal:

1. O princípio da composicionalidade, que estabelece que o significado das sentenças depende do

significado das palavras que as compõem – ou seja, o significado do todo é função do significado das partes e da combinação sintática entre elas. Para saber o significado de uma sentença, é necessário conhecer o significado das suas partes, bem como as regras que definem sua combinação.

2. A condição de verdade, que determina as condições em que tal sentença é verdadeira. Nesse contexto, saber o significado de uma sentença equivale a conhecer suas condições verdade, o que não é o mesmo que saber o seu valor verdade, ou seja, se o fato é verdadeiro ou falso.

3. Os modelos em semântica, em que são construídos sistemas simples, em relação aos sistemas complexos que se desejam estudar. Constrói-se uma teoria lógica para o modelo e, se os resultados forem razoavelmente positivos em relação ao sistema complexo, diz-se que o sistema simples é um bom modelo; caso contrário, ele é abandonado.

A semântica formal considera ainda o fato de que as línguas naturais são utilizadas para falar sobre objetos, indivíduos, fatos, eventos e propriedades, descritos como externos à própria língua – assim, a referencialidade é um de seus aspectos fundamentais. Por essa razão, na semântica formal o significado é entendido, por um lado, como uma relação com a linguagem, e, por outro, com aquilo sobre o que a linguagem fala.

A semântica formal procura responder às seguintes perguntas: o que representam ou denotam as expressões linguísticas?, como calculamos o significado de expressões complexas a partir dos significados de suas partes?

3. RECUPERAÇÃO SEMÂNTICA DA INFORMAÇÃO

Os sistemas de recuperação de informação (SRI) procuram representar o conteúdo dos documentos e apresentá-los ao usuário de maneira a atender rapidamente sua necessidade de informação.

Para tanto, a RI pesquisa técnicas para tratamento, organização e busca de conteúdos a partir do uso de padrões. As ferramentas de RI, geralmente, trabalham com técnicas de indexação capazes de indicar e acessar rapidamente documentos de um banco de dados textual.

Existem três tipos principais de indexação:

- indexação tradicional, em que se determinam os termos descritivos ou caracterizadores dos documentos;

- indexação *full-text* (ou indexação do texto todo), em que todos os termos que compõem o documento fazem parte do índice;

- indexação por *tags* (por partes do texto), em que apenas algumas partes do texto são escolhidas para gerar as entradas no índice (somente as consideradas mais importantes ou mais caracterizadoras).

As buscas são geralmente feitas através de termos fornecidos pelo usuário ou escolhidos por ele entre alguns apresentados. Esses termos podem significar o assunto ou classe a que pertencem os documentos desejados (na indexação tradicional) ou os termos que devem estar presentes nos documentos desejados (nas indexações *full-text* e por *tags*).

Nos sistemas convencionais de busca, as técnicas utilizadas são de base sintática. Entretanto, quando a busca do usuário envolve informação cuja relevância não pode ser dada por palavras-chave, esse modelo não satisfaz. Observa-se, portanto, a crescente exploração das informações semânticas, o que possibilita compreensão dos conceitos em seu contexto e finalidade.

A marcação semântica dos dados na origem é um exemplo das novas tecnologias utilizadas para a RI. A WS tem-se utilizado dessa estratégia para implementar padrões de metadados que adicionem aos dados informações significativas sobre seus contextos, marcando-os semanticamente.

A exploração da semântica intrínseca dos dados busca fundamentos da linguística e da ciência da informação para ampliar o universo de informações recuperadas e a aferição de contextos, por meio do uso de estruturas da linguagem natural, como os sintagmas verbais e nominais, e de ferramentas de representação de relacionamentos semânticos e conceituais.

Técnicas de processamento de linguagem natural (PLN) que permitam a construção de algoritmos para a busca de informação relevante em grande quantidade de documentos estão em crescente demanda. Métodos eficientes de PLN devem fundamentar-se em conhecimentos básicos sobre propriedades da linguagem e principalmente sobre a semântica dos conceitos. Desse contexto emerge a ideia de memória semântica, tema que tem sido objeto de teorias psicolinguísticas e é uma fonte rica para desenvolvimento de modelos computacionais que pretendem se aproximar dos processos mentais usados pelo cérebro humano na compreensão da linguagem. Além dos algoritmos, os modelos computacionais necessitam de dados que representam o conhecimento sobre a linguagem e as associações de senso comum entre os conceitos lexicais e suas propriedades. Atualmente, isso é muito difícil, pois essa tarefa somente pode ser produzida e verificada por humanos. A memória semântica trabalha com um léxico mental – ou seja, conceitos e unidades de conhecimento – e contém informações sobre as relações entre conceitos, formando uma rede conceitual de elementos conectados uns com outros por diferentes tipos de associações.

As redes neurais, por sua vez, são uma representação que têm muitas características comuns à memória humana: podem lidar com informação incompleta ou distorcida, permitem generalizações automáticas e exibem conteúdo baseado no contexto. Essas funções possibilitam várias aplicações na RI guardada na memória humana, essenciais em contextos precisos, e são essas aplicações que se pretendem reproduzir no ambiente computacional.

Um componente semântico que impacta o desempenho de um SRI é o tesauro, ferramenta da linguagem artificial de um domínio conhecido, construído por especialistas para representar através de conceitos o conteúdo informacional, especificando as relações entre eles. Trata-se de um sistema de conceitos que se relacionam entre si e são representados por termos.

Cada termo tem obrigatoriamente vinculação com outro, e é essa vinculação que forma a estrutura do tesauro. Os termos são utilizados pelos indexadores no momento da indexação e devem ser disponibilizados ao usuário no momento da RI.

Silveira e Ribeiro-Neto (2004) estudam o uso automático de tesauro para melhorar resultados de buscas na web, por meio de ranking baseado em conceitos que tem sido estudado para a RI em domínios específicos. Em experimento realizado pelos autores, os termos de busca utilizados em um SRI na web foram comparados aos conceitos de um tesauro, utilizado para encontrar conceitos relacionados. Cada conceito relacionado foi interpretado de forma independente e processado separadamente, e então combinado em uma rede bayesiana[1], para permitir um

ranking final, baseado em conceitos. O objetivo era verificar se a informação acrescida de conceito aumentaria a média da precisão dos resultados da busca.

No estudo foram utilizadas seis fontes de informação: palavra chave, conceito, termo específico, termo geral, termo relacionado e sinônimo. Os autores verificaram, entre outras coisas, que a utilização de um tesauro específico para um domínio específico é fundamental para a melhoria do desempenho das buscas.

O experimento demonstrou o aumento de 30% na média da precisão dos resultados das buscas e, portanto, propõe a visão de que os tesauros melhoram tanto a revocação como a precisão em um SRI.

4. TESAURO

De acordo com Moreira, Alvarenga e Oliveira (2004), o termo "*tesauros*" se origina do grego thesaurós e significa tesouro ou repositório. Esse termo surgiu com a publicação do dicionário analógico de Peter Mark Roget, em Londres, em 1852, intitulado *Thesaurus of English words and phrases*. O termo também designa vocabulário, dicionário ou léxico, porém o dicionário de Roget se diferenciava dos outros por ser um vocabulário organizado de acordo com seu significado, e não por ordem alfabética. A obra teve o mérito de estabelecer a denominação para vocabulários que relacionam seus termos por meio de algum tipo de relação de significado.

Em sua introdução, Roget define seu dicionário como uma "classificação de ideias" e explica que, diferentemente dos outros, o seu permite que se chegue à palavra mais adequada ou à que melhor se ajuste às necessidades do escritor, sem que, de início, ele saiba qual é ela (GOMES, 1996).

Nos anos 1960, o cientista da informação Brian Campbell Vickery apresentou quatro significados para o termo "tesauro" na literatura de sua área, sendo que o significado mais comum é o de uma lista alfabética de palavras, em que cada palavra é seguida de outras relacionadas a ela (VICKERY, 1980 apud MOREIRA; ALVARENGA; OLIVEIRA, 2004).

Currás (1995, p. 88) define tesauro como "uma linguagem especializada, normalizada, pós-coordenada, usada com fins documentários, onde os elementos linguísticos que a compõem – termos, simples ou compostos – encontram-se relacionados entre si sintática e semanticamente".

Tristão (2004, p. 167) o define como "vocabulário de termos, que nada mais é do que uma seleção de termos, baseados em análise de conceitos, na qual se define o termo geral, de maior abrangência, e sua relação com termos mais específicos, que representam os conceitos menores". A Organização Nacional de Padrões de Informação especifica:

> Um tesauro é um vocabulário controlado organizado em uma ordem preestabelecida e estruturado de modo que os relacionamentos de equivalência, de homografia, de hierarquia, e de associação entre termos sejam indicados claramente e identificados por indicadores de relacionamento padronizados empregados reciprocamente. As finalidades primordiais de um te-

> sauro são (a) facilitar a recuperação dos documentos e (b) alcançar a consistência na indexação dos documentos escritos ou registrados de outra forma e outros tipos, principalmente para sistemas de armazenamento e de recuperação de informação pós-coordenados. (ANSI/NISO Z39.19, 2003 apud SALES; CAFÉ, 2008).

Conforme Campos e Gomes (2006), a evolução na construção de tesauros baseia-se em duas vertentes, a americana e a europeia. Os tesauros elaborados nos Estados Unidos a partir da década de 1950 foram fruto do desenvolvimento ocorrido a partir do cabeçalho de assuntos para o unitermo, passando de um sistema pré coordenado para sistemas pós-coordenados.

Silva (2008) afirma que na mesma época, na Inglaterra, o Classification Research Group (CRG) – a partir da Teoria da Classificação Facetada, desenvolvida pelo matemático e bibliotecário indiano Shiyali Ramamrita Ranganathan – ampliou as categorias de personalidade, matéria, energia, espaço e tempo (PMEST) e desenvolveu diversas tabelas de classificação, dando origem a uma técnica denominada Thesaurofacet, que permitiu melhor posicionamento do conceito no sistema de conceitos em uma dada área de assunto, por meio do uso de suas categorias. Conjuntamente, também embasaram os tesauros terminológicos a Teoria da Classificação Facetada, a Teoria do Conceito e alguns princípios terminológicos. Esses instrumentos têm nas características do conceito um elemento essencial para evidenciar as relações entre os conceitos e seu posicionamento no sistema, além de defini-lo.

A citada Teoria do Conceito, voltada para o referente e originalmente chamada Teoria Analítica do Conceito, foi lançada no final da década de 1970 pela cientista da informação Ingetraut Dahlberg, agregando princípios terminológicos relacionados ao conteúdo conceitual e à sua definição. Para Campos e Gomes (2006), essa é uma teoria consolidada para a determinação do que se entenderia por menor unidade em um tesauro: o conceito representado por um termo. Além disso, Moreira (2003) aponta como inovação o uso de definições para o posicionamento do conceito no sistema.

Bräscher (2010) aponta como função dos tesauros a tradução da linguagem dos documentos, dos indexadores e dos pesquisadores a uma linguagem controlada, usada na indexação e na RI em sistemas de informação. Conforme Sales e Café (2008), o ANSI/NISO Z39.19, de 2003, ressalta que os tesauros não são utilizados somente pelos especialistas da informação no momento da indexação, mas também por usuários da informação no momento da busca de documentos.

Segundo Café e Bräscher (2011), nos tesauros as "relações semânticas são estabelecidas por meio da análise das características ou propriedades dos conceitos, as quais permitem identificar diferenças e semelhanças que evidenciam determinados tipos de relacionamentos". Um termo presente em um tesauro pode ser caracterizado de maneiras diferentes dependendo do assunto em questão e também do tipo de sistema que se deseja construir. A estrutura dos tesauros compreende três tipos principais de relações semânticas para relacionar os termos: a hierarquia, a equivalência e a associação.

Nas relações hierárquicas os termos são organizados em gênero/espécie. As relações de equivalência são de sinonímia, ou seja, há termos sinônimos presentes no tesauro e deve-se indicar qual o termo adequado para representar

determinado conceito. As relações de associação, por sua vez, apresentam associações entre os termos, sem especificar qual tipo de relação propriamente existe – são apenas termos que se relacionam de alguma forma.

Nos tesauros são tratados ainda casos de ambiguidade (possibilidade de que uma comunicação linguística se preste a mais de uma interpretação) e de polissemia (possibilidade de que uma palavra comporte mais de um significado).

5. VOCABULÁRIO DE CONTROLE EXTERNO

O tesauro do TCU, denominado Vocabulário de Controle Externo (VCE), foi lançado em 2015 e objetiva ser instrumento de controle terminológico que possibilite padronização no tratamento técnico e maior precisão na recuperação dos conteúdos presentes nos sistemas de informação do TCU.

O inter-relacionamento dos conceitos no VCE foi expresso por meio de relações de três tipos: de equivalência, hierárquicas e associativas. O objetivo das relações é apresentar os descritores em seu contexto semântico.

- **Relação de equivalência**: se termos sinônimos ou quase sinônimos são considerados, para efeito do vocabulário, como representando um mesmo conceito, um deles é escolhido como descritor e os demais são proibidos.

- **Relação hierárquica**: relacionamento que exprime os graus ou os níveis de superordenação e de subordinação dos termos; o termo superordenado representa o gênero de que o termo subordinado é tipo ou espécie.

- **Relação associativa**: reunião de conceitos afins que merecem estar relacionados, mas que não estão ligados por relacionamentos de equivalência nem de hierarquia.

No VCE, cada termo corresponde a um conceito, e todos os termos possuem relacionamentos, sendo o relacionamento determinado pelo significado do termo. As relações entre termos ajudam a compreender os conceitos específicos de controle externo e de áreas correlatas que compõem o tesauro.

Bem além de uma lista hierarquizada de palavras, o VCE é composto de três partes distintas, mas inter-relacionadas. A primeira é formada por palavras chave, relativas às áreas de atuação do Tribunal, acompanhadas por definições e sinônimos; a segunda, correspondente à clientela do TCU e às Entidades de Fiscalização Superiores (EFS) associadas à Organização Internacional de Entidades Fiscalizadoras Superiores (Intosai), traz informações como histórico, nomes alternativos, CNPJ e instituições afins. A terceira parte contempla a toponímia nacional formada pelas regiões, mesorregiões, unidades federativas e municípios brasileiros.

6. APLICAÇÕES DO VCE

6.1 E-JURIS

O e-Juris é uma ferramenta corporativa que faz parte do e-TCU, o qual agrega todos os sistemas de instrução e controle de processos do Tribunal, contendo a mesma lógica, estrutura e apresentação dos demais sistemas de processos. Além disso, é integrado com o portal do TCU, com o sistema corporativo de busca e com outros sistemas do TCU, como o Sagas e o VCE. As premissas adotadas para o novo sistema foram seletividade, qualidade, relevância, tempestividade e simplicidade.

A principal finalidade do e-Juris é a divulgação das teses relevantes, do ponto de vista jurisprudencial, que fundamentaram acórdãos do TCU, por meio de publicações periódicas (boletins e informativos) e de formação e disponibilização, para pesquisas e consultas, da base de dados da jurisprudência do Tribunal. Com isso, o novo sistema unifica processos de trabalho que antes eram feitos separadamente e sem integração.

As teses jurisprudenciais relevantes são representadas por enunciados, sob formato de ementa. Os enunciados representam precedentes jurisprudenciais, não o "entendimento" ou a jurisprudência prevalecente do Tribunal sobre determinada questão.

A adoção da indexação pelo e-Juris permite maior precisão na recuperação dos enunciados. Além da previsão de busca por termos sinônimos, o VCE agrega ao sistema a funcionalidade de sugestão de assuntos correlatos a ser pesquisados, uma vez que a ferramenta é estruturada em um sistema de conceitos inter-relacionados por hierarquia, equivalência e associação.

6.2 BIBLIOTECA DIGITAL DO PORTAL DO TCU

Repositórios corporativos de conhecimento prestam-se a disseminar a informação produzida internamente, a permitir o acesso à cultura organizacional e a subsidiar a informação transformada em conhecimento. Sendo mais que um depósito de documentos, os repositórios podem atuar de forma decisiva, dando suporte ao desenvolvimento de novos produtos e serviços e à formação e à capacitação da força de trabalho da organização. Além disso, costumam servir de fonte oficial de informação aos parceiros e colaboradores, amparar as atividades cotidianas e auxiliar o processo decisório.

No TCU, a biblioteca digital é um dos repositórios corporativos de conhecimento de mais fácil acesso e utilização. Desenvolvida para organizar, tratar e disseminar informações que possam gerar novos conhecimentos, a biblioteca digital permite a inserção de material com diversas tipologias documentais. Assim, no mesmo ambiente é possível encontrar livros, trabalhos acadêmicos, apresentações, cartilhas, periódicos, acordos, contratos, expedientes e normativos, além de imagens diversas e outros tipos de recursos. A ferramenta permite que sejam estabelecidos dois níveis de acesso aos conteúdos depositados: é possível permitir o livre acesso a documentos de caráter público, da mesma forma que é possível limitar o

acesso a documentos de cunho interno ou que possuam algum tipo de restrição.

A inserção de documentos no ambiente é realizada descentralizadamente, e existem gestores de conteúdo responsáveis pela aprovação do que ficará disponível no repositório. A biblioteca digital possui também um formulário de entrada de dados que pode ser usado em todo o Tribunal. Esse formulário é composto de metadados controlados e possibilita a descrição do conteúdo, com elementos tais como título, autoria e data. Mais do que isso, requer que os documentos sejam classificados em uma árvore temática e que os assuntos tratados sejam traduzidos por palavras-chave oriundas de um vocabulário controlado.

Em outras palavras, a biblioteca digital do TCU constitui-se num repositório corporativo de conhecimento que exige que as informações sejam classificadas e indexadas. Para a indexação dos conteúdos, o ambiente está integrado ao VCE. Além disso, como a biblioteca também está integrada à ferramenta de pesquisa textual do portal corporativo, é possível fazer buscas diretamente no portal e recuperar o conteúdo depositado no ambiente dessa biblioteca.

6.3 SISTEMA ORIENTAR

O Sistema Orientar foi concebido para ser ferramenta de orientação, gestão e disseminação do conhecimento sobre controle externo. Ele permite que qualquer servidor do Tribunal possa encaminhar perguntas sobre temas pré-selecionados de controle externo, tais como auditoria, planejamento, contas anuais, tomada de contas especial, representação, denúncia, solicitação, avaliação de qualidade, cobrança executiva, normas de controle externo e demais procedimentos processuais, além de tirar dúvidas sobre o sistema Fiscalis.

Após a seleção do tema, o sistema direciona a pergunta para a unidade responsável pela área. A partir da coletânea das perguntas e respostas, cada unidade respondente cria automaticamente um banco de dados de perguntas mais frequentes (FAQ), que fica armazenado no sistema e disponível para consultas e pesquisas por todos os servidores.

Diversas unidades do TCU já foram cadastradas como unidades respondentes, e o sistema é integrado ao VCE. A adoção do vocabulário controlado permite maior precisão na recuperação da informação desejada, uma vez que são atribuídos, tanto às perguntas quanto às respostas, termos de indexação que representam os assuntos tratados.

6.4 WIKI DE CONTROLE EXTERNO

Entre as diversas possibilidades de armazenar o conhecimento de uma instituição, uma em especial merece destaque numa era em que produzimos conhecimento por meio de variados colaboradores e parceiros: as wikis – ferramenta de tecnologia conhecida como "software social" e desenhada com um conjunto de características que permitem a criação e a organização de conhecimento no mundo colaborativo. A utilização das wikis tem se revelado uma solução de baixo custo, com alto grau de eficiência, para o fomento à criação cooperativa de conhecimento dentro das organizações.

O ambiente wiki é a evolução do conceito de *computer supported cooperative work* (CSCW), que surgiu da necessidade de que organizações tenham pessoas trabalhando em lugares físicos diferentes, ao mesmo tempo que precisando alcançar resultados rápidos conjuntamente; ou seja, surgiu para facilitar a comunicação e a produtividade de grupos remotos.

Desde 2009, o TCU utiliza o software livre Mediawiki para gerir uma wiki de controle externo e acesso restrito aos seus servidores. Trata-se de um importante espaço colaborativo de construção de conhecimento que reúne tutoriais informais e verbetes especializados oriundos do VCE.

A Wiki pode ser acessada e editada por todos os servidores; eles agregam informações aos verbetes e tutoriais em tópicos que abrangem norma, legislação e doutrina. A Wiki é importante ferramenta de gestão do conhecimento, na medida em que fornece informações e documentos úteis ao trabalho cotidiano dos auditores, de acordo com sua área de atuação, e se configura como um genuíno ambiente de inteligência coletiva da organização.

6.5 FUTURAS APLICAÇÕES

O VCE apresenta ainda potenciais aplicações em diversos sistemas de informação do Tribunal. A integração da terminologia na ferramenta de busca do portal do TCU, por meio da adoção do vocabulário controlado tanto para o tratamento quanto para a busca, é fator primordial para o aumento de precisão e rapidez na RI.

Outra possibilidade de uso do VCE é como dicionário em softwares de indexação automática de grandes volumes de documentos, servindo de parâmetro terminológico na área de controle externo. Essa funcionalidade já foi testada durante o processo de indexação automática de enunciados, constante em uma prova de conceito para aquisição de software de *data mining* e análise semântica de dados.

Em um cenário ideal, vislumbra-se que o Tribunal adote padrões, a exemplo de metadados de assunto e da ferramenta de controle terminológico, para garantir a melhoria do desempenho de seus SRI.

NOTAS

1 Redes bayesianas constituem um modelo gráfico que representa simplesmente as relações de causalidade das variáveis de um sistema. Em resumo, redes bayesianas, também conhecidas como redes de opinião, redes causais e gráficos de dependência probabilística, são modelos gráficos para raciocínio (conclusões) baseado na incerteza, em que os nós representam as variáveis (discretas ou contínuas), e os arcos representam a conexão direta entre eles (SILVEIRA; RIBEIRO NETO, 2004).

REFERÊNCIAS

ALMEIDA, M. B.; SOUZA, R. R. Avaliação do espectro semântico de instrumentos para organização da informação. Encontros Bibli – Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 16, n. 31, p. 25 50, 2011. Disponível em: <http://mba.eci.ufmg.br/downloads/11963-60907-1-PB.pdf>. Acesso em: 23 nov. 2016.

ALTOUNIAN, M.; ZAULI, A. A semântica na recuperação da informação na web: novas tendências. 2013. Trabalho apresentado como requisito parcial para aprovação na disciplina Recuperação da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

BRÄSCHER, M. Elaboração de tesauros. Brasília, 2010.

CAFÉ, L.; BRÄSCHER, M. Organização do conhecimento: teorias semânticas como base para estudo e representação de conceitos. Informação & Informação, Londrina, v. 16, n. 3, p. 25 51, jan./jun. 2011.

CAMPOS, M. L. A.; GOMES, H. E. Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador. Perspectivas em Ciência da Informação, Belo Horizonte, v. 11, n. 3, p. 348 359, set./dez. 2006.

CAPRI, D.; GARRIDO, I.; DUARTE, R. Recuperação semântica da informação. 2009. Trabalho apresentado como requisito parcial para aprovação na disciplina Recuperação da Informação, Centro de Ciências da Educação, Universidade Federal de Santa Catarina, Florianópolis, 2009. Disponível em: <http://pt.slideshare.net/doritchka/angel-recuperao-semntica-da-informao>. Acesso em: 23 nov. 2016.

CURRÁS, E. Tesauros: linguagens terminológicas. Brasília, DF: CNPq; Ibict, 1995.

GOMES, H.E. Classificação, tesauro e terminologia; fundamentos comuns. Rio de Janeiro: UNIRIO, 1996.

MOREIRA, A. Tesauros e ontologias: estudo de definições presentes na literatura das áreas das ciências da computação e da informação, utilizando-se o método analítico-sintético. Perspectivas em Ciência da Informação, Belo Horizonte, v. 8, n. 2, p. 216-226, jul/dez. 2003.

_____; ALVARENGA, L.; OLIVEIRA, A. P. Thesaurus and ontology: a study of the definitions found in the Computer and Information Science Literature, by means of an analytical-synthetic method. Knowledge Organization, v. 31, n. 4, p. 231-244, 2004.

SALES, R., CAFÉ, Lígia. Semelhanças e Diferenças entre Tesauros e Ontologias. DataGramaZero - Revista de Ciência da Informação, v.9 n.4 Ago 2008.

SILVA, A. Análise das relações semânticas em tesauros jurídicos brasileiros: orientações das normas e aplicação prática. 2013. Trabalho de Conclusão de Curso (Bacharel em Biblioteconomia) – Centro de Ciências da Educação, Universidade Federal de Santa Catarina, Florianópolis, 2013. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/103801/TCC_Aline_da_Silva_20131PDFA.pdf?sequence=1>. Acesso em: 23 nov. 2016.

SILVA, D., SOUZA, R., ALMEIDA, M.B. Ontologias e vocabulários controlados: comparação de metodologias para construção. Ciência da Informação, v. 37, n. 3, p. 60-75, set/dez. 2008.

SILVEIRA, M. L.; RIBEIRO-NETO, B. Concept-based ranking: a case study in the juridical domain. Information Processing & Management, Doha, v. 4, n. 5, p. 791 805, Sept. 2004.

SOUZA, A. et al. Recuperação semântica de objetos de aprendizagem: uma abordagem baseada em tesauros de propósito genérico. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 19., 2008, Uberlândia. Anais... Uberlândia: SBIE, 2008. p. 603 612.

SOUZA, A.; ROCHA, R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. Perspectivas em Ciência da Informação, Belo Horizonte, v. 11, n. 2, p. 161 173, maio/ago. 2006. Disponível em: <http://www.scielo.br/pdf/pci/v11n2/v11n2a02.pdf>. Acesso em: 23 nov. 2016.

SZYMANSKI, J.; DUCH, W. Information retrieval with semantic memory model. Cognitive Systems Research, Kalamazoo, v. 14, n. 1, p. 84 100, Apr. 2012. Disponível em: <http://ethologie.unige.ch/etho5.10/themes/semantic.memory/szymanski.duch.2011.information.retrieval.in.sematic.memory.neural.network.models.pdf>. Acesso em: 23 nov. 2016.