Versão preprint dos autores (com ajuste do título e poucas correções ortográficas)

As nuvens de termos aplicadas à análise da pós-graduação interdisciplinar

Vinícius Medina Kern, Alexandre Leopoldo Gonçalves, Alessandro Botelho Bovo

INTRODUÇÃO

Este capítulo tem objetivo duplo: (i) apresentar a descoberta de conhecimento em texto ou KDT (do inglês *knowledge discovery in text*) como técnica da engenharia do conhecimento para análise e apoio à tomada de decisão na pós-graduação interdisciplinar e (ii) explicar como foram compostas as "nuvens de termos" (*word clouds*, daqui em diante referidas sem aspas) que ilustram cada capítulo deste livro, construídas com o software Wordle (FEINBERG, 2009). A composição das nuvens requer a **extração de termos** presentes no texto (incluindo sua identificação, contagem e organização), etapa inicial de KDT, além da aplicação de um algoritmo de leiaute gráfico.

A **engenharia do conhecimento**, numa definição clássica, é "uma disciplina que envolve integrar conhecimento em sistemas computacionais para resolver problemas complexos que usualmente requereriam um alto nível de perícia humana" (FEIGENBAUM; McCORDUCK, 1983). Schreiber et al. (2000) a apresentam como apoiadora da gestão do conhecimento, em especial na externalização – a conversão de conhecimento tácito para explícito. KDT, KDD (*knowledge discovery in databases*) e ontologias (representações de conceitos de determinado domínio) são alguns dos tópicos em evidência na engenharia do conhecimento.

A etapa de extração de termos em KDT usualmente envolve a construção de vetores multidimensionais nos quais as dimensões são os termos ocorrentes e as coordenadas

correspondem às frequências (contagens) desses termos no texto[1]. Ou seja, cada capítulo é um vetor. Sua exibição na forma de nuvem de termos é uma forma peculiar[2] de representação.

Além de peculiar, a representação de textos como nuvens de termos é útil, pois uma representação vetorial como a da Figura 1 (em 2 dimensões) seria precária em 3 dimensões e inviável em mais do que 3. Numa base vetorial na qual cada termo usado neste livro é uma dimensão, os vetores correspondentes aos capítulos ocorrem na forma { (termo1, freq1); (termo2, freq2); ...; (termo***n***, freq***n***) }. Na inviabilidade da representação vetorial gráfica além da terceira dimensão, a nuvem de termos é uma solução pragmática.

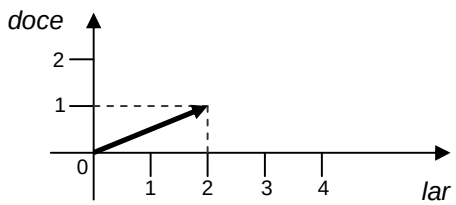

Figura 1. Representação vetorial do adágio “Lar, doce lar”, no qual as dimensões são “lar” e “doce”

Representáveis graficamente ou não, os vetores podem ser operados de forma algébrica. Soma, produto, distância escalar e angular e outras aplicações mais sofisticadas – a tudo estão sujeitos os capítulos (ou melhor, seus vetores). Essa vetorização de textos, além de aplicável à

[1] Quando há vários documentos-fonte para a vetorização, a frequência de cada termo é ponderada, ou seja, não é usada apenas a contagem total de cada termo. Essa ponderação é feita usualmente por meio do peso TF*IDF (*term frequency * inverse document frequency*).

[2] “Peculiar” porque não há uma maneira “correta” de distribuir os termos no espaço. Há regras claras para determinar o tamanho das letras e para garantir a não-sobreposição, mas não há regra fixa para posicionar cada termo específico.

construção de nuvens de termos, é ponto de partida de um processo de inferência e aplicação de algoritmos especializados conhecido como mineração de textos ou *text mining*.

Gonçalves (2006) observa que essa inferência ou descobrimento de padrões em textos em linguagem natural pode revelar conhecimento útil, aplicável à tomada de decisão. A expressão "descoberta de conhecimento em texto" ou KDT surge dessa potencialidade[3].

A seguir, explicamos como foram compostas as nuvens de termos dos capítulos. Depois, apresentamos um breve panorama sobre KDT como forma de revelar o potencial dessa tecnologia, aqui "encoberto pelas 'nuvens'", para construir análises na pós-graduação interdisciplinar. Finalmente, ilustramos esse tipo de análise com um cenário de aplicação.

NUVENS DE TERMOS DOS CAPÍTULOS DESTE LIVRO

Descrevemos brevemente a seguir o software usado para criar as nuvens e detalhamos as escolhas de leiaute que levaram às figuras ilustrativas dos capítulos.

O Software Wordle

Wordle é um aplicativo na web autodefinido como "um brinquedo para gerar 'nuvens de palavras'" (FEINBERG, 2009) em arranjo gráfico em duas dimensões. O tamanho de cada palavra tem relação com a frequência no texto e há opções variadas de tipo (*font*), direção do texto e cores. Palavras comuns (*stopwords*) como artigos e algumas preposições são eliminadas, porém em apenas uma língua (mas é possível eliminar manualmente outras *stopwords* ou quaisquer termos que se desejar).

[3] Cabe assinalar a distinção feita por Fayyad et al. (1996), segundo a qual KDD (o mesmo vale para KDT) denota o processo de extração de conhecimento, ao passo que *data mining* (idem para *text mining*) refere-se à fase de aplicação de técnicas e algoritmos especializados. Neste texto, usamos KDT e *text mining* de forma intercambiável, da forma semelhante a outros autores que não enfatizam essa distinção.

O aplicativo foi escrito como projeto pessoal, embora incorpore (com permissão) código desenvolvido pelo autor para a IBM Research. A página web do Wordle não dá maiores detalhes sobre algoritmos e técnicas computacionais usadas, mas é razoável supor que o programa para coletar os termos e frequências seja parte de sistemas de *text mining*.

O texto fonte para construir a nuvem pode vir de um *blog* ou outro sítio que use *web feeds*, artefato comum em sítios que distribuem conteúdo para agregação posterior (serviço anunciado usualmente por meio do ícone da Figura 2). Pode-se, também, colar texto no próprio sítio do Wordle para gerar as nuvens de termos.

Figura 2. Ícone usual em um sítio que oferece web feeds. Fonte: http://en.wikipedia.org/wiki/Web_feed

Há uma série de restrições à apropriação da tecnologia, mas o uso das imagens geradas é livre sob uma licença Creative Commons[4]. O programa roda no computador do usuário. Atualmente, existem aplicativos similares na web tais como TagCrowd (tagcrowd.com), Tagul (tagul.com) e WordItOut (worditout.com).

Como foi Criada Cada Nuvem

As nuvens de termos de cada capítulo deste livro foram criadas a partir da cópia-e-cola do texto do capítulo na janela apropriada no Wordle[5]. A distribuição dos termos no plano é consequência da aplicação de um algoritmo (não explicitado pelo autor) do Wordle. É possível

[4] http://creativecommons.org/licenses/by/3.0/us/

[5] http://www.wordle.net/create

redistribuir os termos, gerando um conjunto aparentemente ilimitado de leiautes. As características de leiaute escolhidas foram:

- Tipo (*font*): **ExpresswayFree**. A escolha seguiu um critério estético, privilegiando a boa legibilidade dos termos.
- Cores: **BW**. Os termos estão em preto sobre fundo branco (para impressão simples em papel claro).
- Idioma para remoção de *stopwords*: **Português**. Palavras estrangeiras comuns (e.g., "*et*" e "*al*." do latim, "*and*", "*the*" e "*of*" do inglês, "*dans*" e "*les*" do francês), existentes pelo menos nas listas de referências, foram eliminadas individualmente na nuvem gerada (com clique direito sobre o termo e opção por remover).
- Distinção maiúsculo-minúsculo (*case sensitiveness*) no menu "Language": **Minúsculas** (*Make all words lower-case*). Isso evita apresentar como termos distintos, por exemplo, "pesquisa" e "Pesquisa", embora imponha minúsculas para nomes próprios e siglas.
- *Stemming*: O Wordle não equaciona palavras semelhantes com mesmo radical (e.g., "disciplina" e "disciplinar"), processo conhecido como *stemming*. As únicas simplificações feitas no processo são a eliminação de palavras comuns (*stopwords*) e a redução a minúsculas.
- Disposição das palavras no menu "Layout": **Vertical** (todas). Isso permitiu criar uma imagem legível com dimensões compatíveis com uma página de livro.

Essa especificação permite a qualquer pessoa reeditar a construção da nuvem de termos de um capítulo, embora o arranjo espacial – o leiaute gráfico – possa ser distinto. É possível

refazer o leiaute gráfico mantendo as demais características. Os termos exibidos, entretanto, serão sempre os mesmos.

O Que A Nuvem nos Diz

O primeiro contato com a nuvem costuma gerar certa fascinação acompanhada de tentativas de interpretação. Não é de surpreender – afinal, até um computador consegue tirar conclusões a partir de nuvens de termos. É possível usar técnicas automáticas para desambiguação – para esclarecer qual o gene em questão dentre os vários que compartilham a mesma sigla, ou para saber se "ontologia" se refere à filosofia ou ao artefato tecnológico (e ainda se no contexto da Ciência da Informação ou da Ciência da Computação).

Se até máquinas interpretam, com maior razão o cérebro humano pode buscar sentido nos termos destacados e suas possíveis conexões. Num sentido estrito, porém, é escassa a possibilidade de tirar conclusões a partir da nuvem de termos feita no Wordle.

Pode-se concluir com precisão que os termos mostrados correspondem aos utilizados e suas frequências, exceto *stopwords*. Porém, mesmo essa interpretação é limitada pela ausência de *stemming* (que faz com que "interdisciplinar" e "interdisciplinares" sejam tratados como termos absolutamente independentes), pelo não-tratamento de sinônimos (perdendo a conexão entre "velho" e "idoso", por exemplo) ou polissêmicos (que faz com que o uso de "ontologia" tanto no contexto da filosofia quanto no contexto das ciências da informação e da computação resulte num único termo na nuvem, o mesmo que acontece com as ocorrências do substantivo e do verbo "ser"). Linguagem figurada, sentido metafórico... nada disso aparece explicitamente na nuvem.

A nuvem é, então, uma singela representação gráfica do uso de palavras. Fica a cargo do leitor a sua possível apreciação como uma espécie de marca, de *fingerprint* do conteúdo abordado, bem como da harmonia ou desarmonia do arranjo, da prevalência de um tema específico ou da

ênfase na questão da interdisciplinaridade, ou, ainda quem sabe, da revelação da possível interdisciplinaridade tecida no texto.

CO-OCORRÊNCIA, CONTEXTO E APLICAÇÕES DE KDT

A extração de elementos textuais para aplicações de KDT é obtida usualmente por meio de técnicas como processamento de linguagem natural, estatística e recuperação de informação (GONÇALVES, 2006). Os procedimentos específicos na fase inicial do processo de KDT incluem a identificação de termos, a eliminação de *stopwords* (palavras comuns, como alguns artigos e preposições) e podem incluir sofisticações como *stemming* (a reunião de termos com radical comum – por exemplo, "interdisciplina" e "interdisciplinar"), entre outros.

A **co-ocorrência** de termos em um mesmo texto é o fundamento de muitas abordagens de KDT. Maher e Simoff (1997) consideravam-na a base de virtualmente tudo, porém Cohen e Hunter (2008), mais recentemente, apontam outras duas abordagens – os sistemas baseados em regras e os sistemas baseados em aprendizagem de máquina.

Co-ocorrências Criam Contexto

Essa co-ocorrência pode ser ilustrada com os resultados de uma busca por "*knowledge discovery in text*" no Portal Inovação em 19/04/2010. Entre os currículos Lattes[6] que citam a expressão no título, palavras-chave ou "Outras informações" nos itens da produção intelectual,

[6] Na verdade, não apenas entre os currículos Lattes, mas também entre currículos sumários de especialistas depositados diretamente no Portal Inovação, embora numa escala muito menor. Em 19 de abril de 2010, para cada currículo sumário existiam cerca de 3 mil currículos Lattes – mais precisamente, havia 613 currículos sumários e 1.793.596 Lattes (Fonte: função "Cartograma" do Portal Inovação, http://www.portalinovacao.mct.gov.br/).

os 20 termos co-ocorrentes (i.e., que constam nos mesmos itens) mais frequentes são os listados na Figura 3.

Ontologias	Competitividade	Economia	Banco de dados
Sistemas de informação	Qualidade	Brasil	Descoberta de conhecimento
Data Warehouse	Carne Suína	Crédito Agrícola	Inovação
Gestão	Engenharia do conhecimento	Gestão do conhecimento	Internet
Text Mining	aiNet	ontoKEM	Plataforma Lattes

Figura 3. Termos co-ocorrentes com "*knowledge discovery in text*" em itens de produção científica, tecnológica ou artística em currículos no Portal Inovação

Entre os termos co-ocorrentes há sinônimos e correlatos como “Text Mining” e “Descoberta de conhecimento”, um algoritmo para análise de dados como “aiNet” (CASTRO; VON ZUBEN, 2001), um software para a elaboração de ontologias computacionais como “ontoKEM” (RAUTENBERG; TODESCO; GAUTHIER, 2009), a disciplina em que se enquadra KDT (“Engenharia do conhecimento”), especialidades relacionadas como “Sistemas de informação” e “Data Warehouse”, áreas e ambientes nos quais KDT se aplica etc. Não há garantia de que os termos co-ocorrentes sejam ou incluam sinônimos, super- ou subáreas ou qualquer outro tipo de termo. O que se pode afirmar é que as co-ocorrências estabelecem certo **contexto** para o termo em questão.

Contexto Possibilita Entendimento e Descoberta

Esse contexto é que permite a gama de aplicações de KDT/*text mining* existentes. Fan et al. (2006) caracterizam o papel de *text mining* como o de filtrar vastas coleções de dados

semiestruturados ou não estruturados[7], examinar as fontes de informação, ligar conceitos em documentos distantes, mapear relações e ajudar a responder perguntas.

Van Haagen et al. (2009) reconhecem o pioneirismo de Swanson (1986), que, num trabalho seminal, mostrou por meio de técnicas de *text mining* que o óleo de peixe poderia tratar a síndrome de Raynaud (caracterizada pela descoloração de dedos em resposta ao frio ou estresse emocional). A hipótese baseava-se no fato, registrado na literatura, de que alguns parâmetros fisiológicos anormais na presença da doença eram sabidamente (também segundo registros na literatura) passíveis de normalização por meio da ingesta do óleo de peixe. Nenhum estudo, até então, havia testado se o óleo de peixe poderia melhorar os sintomas da doença. Um teste clínico posterior comprovou a hipótese (SWANSON, 1993).

Técnicas de *text mining* vêm sendo usadas como parte do ferramental metodológico para a **desambiguação** de termos. Por exemplo, o problema do uso de siglas distintas para representar o mesmo gene e o uso da mesma sigla para representar genes distintos foi abordado por Mons (2005), que usou técnicas de *text mining* em conjunto com ontologias para fazer mapeamentos entre uma base de conceitos e os termos que aparecem num texto, relacionados por meio de seus contextos (ou nuvens de termos). Essa técnica permitiria esclarecer, por exemplo, se BSE em “Epidemiological considerations of BSE” se refere a “Breast Self Examination” ou a “Bovine Spongiform Encephalopathy”. Da mesma forma, seria possível reconhecer a proteína CD40 Ligand em qualquer menção a suas muitas designações alternativas – TNFSF5, IMD3, T-BAM, CD 154 etc.

[7] Um texto em linguagem natural é estruturado no sentido de possuir uma estrutura sintática, mas aqui a referência a “estrutura” é feita no âmbito da Computação. Os dados ditos “estruturados” estão em bancos de dados – identificados, indexados e armazenados em registros e campos específicos. Dados “semiestruturados” possuem marcação com *tags* em linguagem XML. Textos em e-mails, relatórios, artigos etc., nesse sentido, são considerados dados “não estruturados”.

A área biomédica é terreno fértil para KDT e impulsiona o progresso na área. Cohen e Hersh (2005) e Zweigenbaum et al. (2007) revisaram a literatura sobre os avanços em *text mining* biomédica e apontaram o **reconhecimento de entidades** (*named entity recognition*), a **classificação de textos**, a **extração de sinônimos e abreviações**, a **identificação de relacionamentos**, a **geração de hipóteses** (como a do tratamento da doença de Raynaud com óleo de peixe), a **sumarização de textos**, a ***question answering*** (resposta automática a perguntas feitas em linguagem natural) e a **descoberta baseada na literatura** (a descoberta de ligações indiretas entre entidades – da qual o achado de Swanson (1986) também é exemplo) entre as áreas de avanço mais promissor.

Dada a existência de um *corpus* textual, há possibilidade de aplicação de KDT/*text mining* segundo a gama de possibilidades citadas. Havendo bases de dados estruturadas, pode-se aplicar KDD/*data mining* – a coleção de técnicas congêneres aplicáveis a bases de dados estruturadas. A próxima seção discute essas possibilidades no contexto da pós-graduação interdisciplinar.

DESCOBERTA DE CONHECIMENTO APLICADA À PÓS-GRADUAÇÃO INTERDISCIPLINAR

A nuvem de termos instiga a percepção e a interpretação subjetivas. Porém, é possível investigar características objetivas de um *corpus* textual usando técnicas de descoberta do conhecimento. Nesta seção, apresentamos um estudo rápido da evolução do uso do termo "interdisciplinaridade" e suas co-ocorrências em currículos Lattes, bem como criamos um cenário de aplicação de técnicas de descoberta do conhecimento (KDT e KDD) na análise da pós-graduação interdisciplinar.

Além da Nuvem: Estudo da Evolução da Ocorrência de "Interdisciplinaridade"

A Plataforma Lattes registra o histórico da produção intelectual brasileira. Embora lançada em agosto de 1999, os currículos depositados relatam a produção anterior, ainda que em menor cobertura. Apresentamos a seguir os resultados de um estudo expedito que buscou responder à pergunta: Como tem evoluído o uso do termo “interdisciplinaridade” e suas co-ocorrências na produção intelectual brasileira?

A Figura 4 mostra o resultado desse estudo realizado pelo segundo e terceiro autores deste capítulo, mostrando a evolução no tempo do uso do termo “interdisciplinaridade” e alguns de seus termos co-ocorrentes mais frequentes em currículos vitae (CVs) na Plataforma Lattes. O estudo foi realizado sobre uma base curricular atualizada até 2008, da qual se extraiu o conjunto de currículos que contêm a cadeia de caracteres “interdisciplinar” nos itens da produção científica, tecnológica e artística, usando funções de mineração de textos da Plataforma ISEKP do Instituto Stela[8]. A escolha da escala logarítmica de frequências se propõe a favorecer a legibilidade de valores de frequência muito grandes (nos anos recentes) e muito pequenos (na era pré-Lattes).

[8] http://www.stela.org.br/

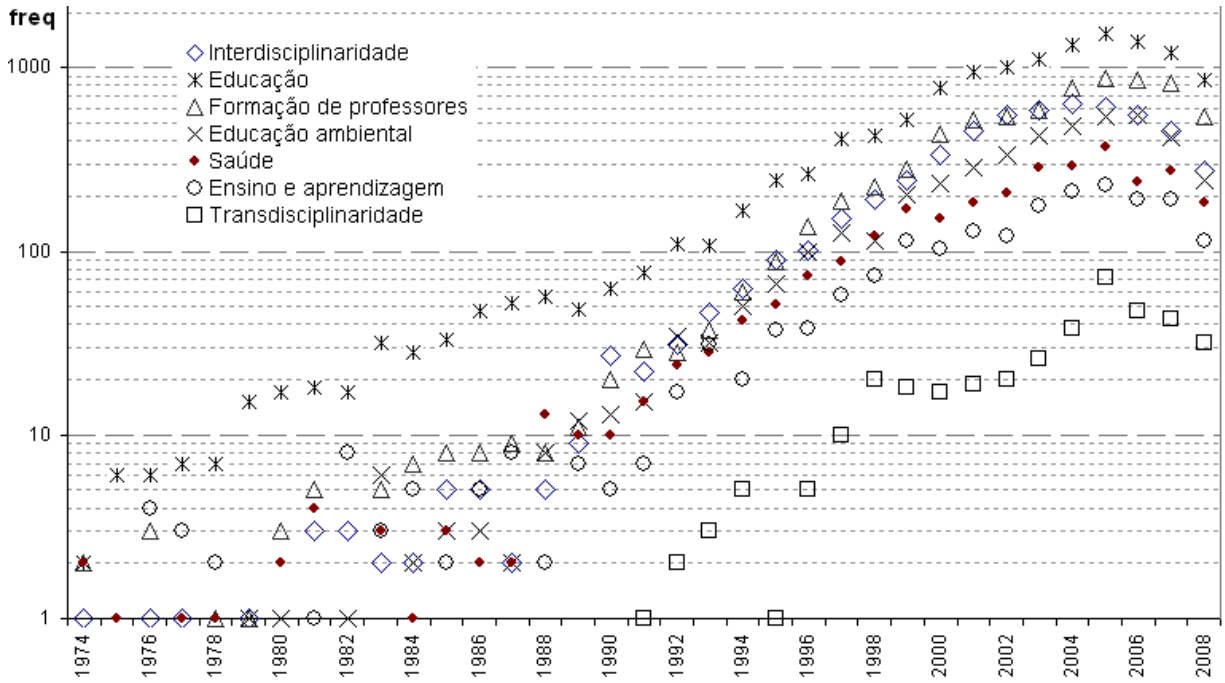

Figura 4. Evolução da frequência de termos co-ocorrentes com "interdisciplinaridade" em CVs Lattes

Dos termos que contêm a cadeia de caracteres “interdisciplinar”, “interdisciplinaridade” é o único com ocorrência histórica relevante nos currículos Lattes. O gráfico registra a frequência com que o termo aparece nos currículos, ano a ano, juntamente com algumas de suas principais co-ocorrências.

A primeira menção a “interdisciplinaridade” é de um item curricular datado de 1974. Pode-se observar que a evolução das frequências dos diversos termos é relativamente harmônica – ou seja, não houve alguma “moda” passageira a vincular “interdisciplinaridade” a algum outro termo. A co-ocorrência com comportamento mais peculiar é “transdisciplinaridade”, que surgiu apenas em 1991, mas a partir daí evoluiu de forma semelhante às demais co-ocorrências. Nota-se um decaimento generalizado nas frequências a partir de 2006, mas isso pode decorrer da defasagem entre a realização de uma publicação e seu registro no CV correspondente.

De forma semelhante ao estudo ilustrado na Figura 4, seria possível investigar outras questões interessantes sobre a marca temporal da interdisciplinaridade na literatura acadêmica, tais como: Quais os pesquisadores e periódicos pioneiros no uso do termo? Houve variação no conjunto de termos co-ocorrentes? Quais as co-ocorrências que, embora pouco frequentes, mostram-se como tendências?

Aplicação: Análises no Âmbito da Pós-Graduação Interdisciplinar

Docentes, discentes e gestores de programas acadêmicos interdisciplinares, bem como gestores institucionais (pró-reitores de pesquisa, em especial), têm contato com uma gama de fontes de informação passíveis de análise por meio de ferramentas de descoberta do conhecimento. A Figura 5 enumera algumas dessas fontes e aponta sua disponibilidade.

	Disponibilidade usual para programas de pós-graduação e instituições acadêmicas		
Fonte de informação	**Dados estruturados** (em bancos de dados)	**Dados semi-estruturados** (usualmente em XML)	**Dados não-estruturados** (usualmente texto)
Teses e dissertações	**Não**: usualmente, apenas metadados são tratados em bancos de dados.	**Não**: usualmente, apenas metadados têm versão XML armazenada.	**Sim**: além das próprias, muitas teses e dissertações estão disponíveis em bancos de teses.
CVs Lattes	**Não**: usualmente, apenas os administradores do sistema-fonte dispõem de dados ou metadados estruturados.	**Sim**: indivíduos podem gerar seus CVs em XML e instituições podem obtê-los via convênio com o CNPq. Outros interessados podem obter CVs em XML via ferramentas gratuitas como scriptLattes (MENA-CHALCO; CESAR JUNIOR, 2009).	**Sim**: os CVs, inclusive de terceiros, estão disponíveis na web, em HMTL.
Bibliotecas digitais de artigos científicos		**Sim**: bibliotecas digitais como o SciELO (www.scielo.br) costumam disponibilizar versões de artigos em XML.	**Sim**
Tabela de áreas do conhecimento (CNPq/Capes)		**Não**, embora seja pouco trabalhoso construir uma versão em XML.	**Sim**
Classificação Qualis (Capes)		**Não**, embora seja pouco trabalhoso construir uma versão em XML.	**Sim**

Figura 5. Algumas fontes de informação relacionadas a programas de pós-graduação e sua disponibilidade

Teses e dissertações são fontes cujas informações costumam permanecer em grande parte encobertas. A aplicação de KDT, no entanto, pode revelar conhecimento interessante[9]. Uma abordagem promissora é o reconhecimento de entidades e a identificação de suas possíveis relações (ZHU et al., 2007) a partir do texto. Essa abordagem pode ser aplicada a bases de teses e dissertações.

O reconhecimento de entidades em teses e dissertações requer, além da extração dos termos do texto, um cruzamento de informações com bases externas. Dessa forma, entre as entidades estariam candidatos, orientadores, examinadores, autores e obras citadas, revistas científicas, eventos, editoras, termos predominantes no texto, instituições dos examinadores externos etc.

Essa abordagem permitiria responder a perguntas de especial interesse para programas interdisciplinares, tais como: Há uma literatura de referência usualmente citada nas teses e dissertações (ou cada candidato cita um conjunto peculiar de autores e obras)? Os termos predominantes na tese têm relação com o objeto de pesquisa do programa? Em que áreas do conhecimento se enquadram as revistas científicas citadas[10]? Quais são os percentuais de fontes da literatura primária citadas em cada tese?

Além dessas questões, um pró-reitor de pesquisa poderia ter interesse nos padrões de interdisciplinaridade de seus programas e perguntar: Quais são as áreas de formação dos examinadores de teses do programa X? Quais são as áreas do conhecimento dos

[9] "Conhecimento interessante" (*interesting knowledge*) é uma expressão consagrada na literatura como o que se pretende revelar com técnicas de descoberta do conhecimento. Por exemplo, Facca e Lanzi (2005) apresentam um estudo sobre a mineração de conhecimento interessante em blogs (weblogs).

[10] A resposta a essa pergunta pressupõe um cruzamento da informação extraída da tese com a classificação determinada pela tabela Qualis da Capes ou outra base que associe revistas a áreas do conhecimento.

examinadores externos do programa Y? Um gestor de agência de fomento ou de formulação de políticas públicas poderia buscar respostas, ainda, para perguntas como: Quais os percentuais de citação de literatura nacional e estrangeira por programa de pós-graduação? Como se caracteriza a participação de examinadores externos (indivíduos e instituições de vínculo) em cada programa de pós-graduação?

Enquanto teses e dissertações oferecem amplas possibilidades de exploração de dados não-estruturados, a informação curricular própria de um indivíduo ou instituição tem a vantagem da disponibilidade em forma semiestruturada. Isso permite construir, a partir de registros curriculares de docentes de um programa interdisciplinar, uma nuvem de termos como a da Figura 6 e perguntar o quanto esses termos correspondem à proposta do curso de pós-graduação.

Figura 6. Nuvem de termos hipotética obtida a partir de registros curriculares de docentes de um programa de pós-graduação interdisciplinar

Os currículos Lattes podem ser fonte, ainda, para a extração das revistas nas quais um grupo docente publica. Um cruzamento de informações com a classificação de revistas Qualis poderia

gerar a Tabela 1, que descreve os números da produção (hipotética) de um grupo docente vinculado a um programa de pós-graduação.

Tabela 1. Números de publicações, classificadas pelo Qualis, de um grupo docente hipotético

Qualis	Ano 1	Ano 2	Ano 3	Ano 4	Ano 5	Ano 6	TOTAL
A1	2	3	1	0	0	2	*7*
A2	0	1	1	2	0	3	*7*
B1	5	1	5	5	8	7	*31*
B2	4	8	7	1	6	3	*29*
B3	2	3	6	6	15	15	*47*
B4	10	6	20	21	18	4	*83*
B5	8	3	6	9	5	4	*35*
Não-qualis	18	23	8	7	5	18	*70*
TOTAL	*49*	*48*	*54*	*51*	*57*	*56*	*309*

Esses números agregados levam a perguntas: Por que o número de itens aumenta em algumas faixas e diminui em outras? (Tem variado a qualidade da produção, ou mudaram os critérios de classificação das revistas?) A produção em cada faixa é distribuída no grupo ou há concentração em poucos docentes?

Esses foram alguns exemplos de análises viáveis a partir da extração de entidades em várias fontes de informação de interesse de programas de pós-graduação. Embora útil para qualquer programa de pós-graduação, procuramos dar ênfase a questões e análises de interesse específico da pós-graduação interdisciplinar.

CONSIDERAÇÕES FINAIS

Neste capítulo, explicamos como foram montadas as nuvens de termos que ilustram cada capítulo do livro. A construção das nuvens requer procedimentos relativamente simples e o

resultado pode ser apreciado como mera curiosidade, sem que seja necessário haver uma interpretação da nuvem gerada.

Ainda assim, a contagem de termos necessária para gerar a nuvem é, também, etapa fundamental de abordagens computacionais mais sofisticadas, como a descoberta de conhecimento em texto ou KDT, técnicas da engenharia do conhecimento que permite construir contexto e, em conjunto com outras técnicas, adicionar semântica e permitir o raciocínio automático. Por isso, o capítulo também apresentou um breve panorama sobre KDT e aventou sua aplicação em análises no âmbito da pós-graduação interdisciplinar.

A tecnologia descrita é atual. Algumas aplicações pontuais estão disponíveis de forma aberta e sem custo – por exemplo, no Portal Inovação, que oferece buscas sobre toda a base curricular brasileira e, no acesso restrito (sem custo, porém sob *login* senha), a possibilidade de compor graficamente redes conforme vários tipos de vínculo (co-autoria, orientação, co-participação em projetos etc.) e gerar indicadores e informações estratégicas.

Para um programa de pós-graduação interdisciplinar, o apoio desse tipo de técnica cria a oportunidade de ampliar a reflexão sobre a qualidade da atuação. Pode ajudar, também, a interpretar a realidade do programa em relação à avaliação recebida da Capes e a tomar decisões visando à sua evolução.

O cenário de aplicação descrito, no qual pessoas fazem perguntas e têm acesso a respostas com o apoio de artefatos (neste caso, de software), caracteriza o que Fuchs (2005) denominou **sistema sociotecnológico**. Nesses sistemas complexos, não é suficiente considerar que um sistema técnico é construído para servir a um sistema social, ou que um sistema social é afetado por um sistema técnico. É necessário compreender que o desempenho do sistema depende da colaboração dinâmica de agentes humanos e artificiais, sendo que esses últimos desempenham tarefas que poderiam ser – e às vezes são – realizadas por pessoas, embora a

um custo frequentemente proibitivo. Dessa forma, a aplicação de KDT à análise da pós-graduação interdisciplinar permite incorporar conhecimento novo, obtido de forma semiautomática e custo-efetiva, à reflexão e à tomada de decisão.

REFERENCIAS

CASTRO, L. N.; VON ZUBEN, F. J. aiNet: An artificial immune network for data analysis. In: ABBASS, H. A.; SARKER, R. A.; NEWTON, C. (Orgs.). **Data mining**: a heuristic approach. Idea Group, 2001, p. 231-259.

COHEN, A. M.; HERSH, W. R. A survey of current work in biomedical text mining. **Briefings in Bioinformatics**, p. 57-71, 2005.

COHEN, K. B.; HUNTER, L. Getting started in text mining. **PLoS Computational Biology**, v. 4, n. 1, e20, 2008.

FACCA, F. M.; LANZI, P. L. Mining interesting knowledge from weblogs: a survey. **Data & Knowledge Engineering**, v. 53, n. 3, p. 225-241, 2005.

FAN, W.; WALLACE, L.; RICH, S.; ZHANG, Z. Tapping the power of text mining. **Communications of the ACM**, v. 49, n. 9, p. 76-82, 2006.

FAYYAD, U. M. Data mining and knowledge discovery: making sense out of data. **IEEE Intelligent Systems**, v. 11, n. 5, p. 20-25, 1996.

FEIGENBAUM, E. A.; McCORDUCK, P. **The fifth generation**, 1st ed. Addison-Wesley, 1983.

FEINBERG, J. Wordle – Beautiful word clouds. 2009. Disponível em: <http://www.wordle.net/>. Acesso em: 01/03/2010.

FUCHS, C. The internet as a self-organizing socio-technological system. **Cybernetics and Human Knowing**, v. 12, n. 3, p. 57-81, 2005.

GONÇALVES, A. L. **Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à engenharia e gestão do conhecimento**. 2006. 196 f. Tese (Doutorado em Engenharia de Produção) - Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2006.

MAHER, M. L.; SIMOFF, S. Knowledge discovery in multimedia design case bases. In: VERMA, B.; YAO, X. (eds.) **Proceedings of ICCIMA'97**, Griffith University, Gold Coast, p. 6-11, 1997.

MENA-CHALCO, J. P.; CESAR JUNIOR, R. M. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v.15, n. 4, p. 31-39, 2009.

MONS, B. Which gene did you mean? BMC Bioinformatics, v. 6, p. 142, 2005.

RAUTENBERG, S.; TODESCO, J. L.; GAUTHIER, F. A. O. Processo de desenvolvimento de ontologias: uma proposta e uma ferramenta. **Revista Tecnologia** (UNIFOR), v. 30, n. 1, p. 133-144, 2009.

SCHREIBER, A. T.; AKKERMANS, H.; ANJEWIERDEN, A.; HOOG, R.; SHADBOLT, N.; VELDE, W.; WIELINGA, B. **Knowledge engineering and management**: the CommonKADS methodology. MIT Press, 2000.

SWANSON, D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. **Perspectives in Biology and Medicine**, v. 30, n. 1, p. 7-18, 1986.

SWANSON, D.R. Intervening in the life cycles of scientific knowledge. **Library Trends**, v. 41, p. 606-631, 1993.

VAN HAAGEN, H.H.H.B.M.; 't HOEN, P.A.C.; BOVO, A.B.; DE MORRÉE, A.; VAN MULLIGEN, E.M.; CHICHESTER, C.; KORS, J.A.; DEN DUNNEN, J.T.; VAN OMMEN, G.-J.B.; VAN DER

MAAREL, S.M.; KERN, V.M.; MONS, B.; SCHUEMIE, M.J.; RUTTENBERG, A. Novel protein-protein interactions inferred from literature context. **Plos One**, v. 4, p. e7894, 2009.

ZHU, J.; GONÇALVES, A. L.; UREN, V.; MOTTA, E.; PACHECO, R. C. S.; SONG, D.; RUGER, S. Community relation discovery by named entities. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND CYBERNETICS, 2007, Hong Kong. **Proceedings of ICMLC**, 2007, v. 4, p. 1966-1973.

ZWEIGENBAUM, P.; DEMNER-FUSHMAN, D.; YU, H.; COHEN, K. B. Frontiers of biomedical text mining: current progress. **Briefings in Bioinformatics**. v. 8. n. 5, p. 358-375, 2007.