Workflows Científicos com Apoio de Bases de Conhecimento em Tempo Real

Victor S. Bursztyn[1,2], Jonas Dias[2], Marta Mattoso[1]

[1]Programa de Engenharia de Sistemas e Computação/COPPE – Universidade Federal do Rio de Janeiro (UFRJ), Brasil

[2]EMC Brazil Research & Development Center – Parque Tecnológico, Ilha do Fundão, Rio de Janeiro, RJ, Brasil

{vbursztyn,marta}@cos.ufrj.br

jonas.dias@emc.com

Abstract. *One major challenge in large-scale experiments is the analytical capacity to contrast ongoing results with domain knowledge. We approach this challenge by constructing a domain-specific knowledge base, which is queried during workflow execution. We introduce K-Chiron, an integrated solution that combines a state-of-the-art automatic knowledge base construction (KBC) system to Chiron, a well-established workflow engine. In this work we experiment in the context of Political Sciences to show how KBC may be used to improve human-in-the-loop (HIL) support in scientific experiments. While HIL in traditional domain expert supervision is done offline, in K-Chiron it is done online, i.e. at runtime. We achieve results in less laborious ways, to the point of enabling a breed of experiments that could be unfeasible with traditional HIL. Finally, we show how provenance data could be leveraged with KBC to enable further experimentation in more dynamic settings.*

1. Introdução

Devido a um maior acesso a recursos do Processamento de Alto Desempenho (PAD), tanto o mundo científico quanto o corporativo têm visto uma ascensão das aplicações com intensa manipulação de dados. Nesse contexto, a modelagem por meio de *workflows* surge como uma maneira apropriada para capturar os passos do fluxo de transformações de dados (ou *dataflow*) de tais aplicações. Na gerência da execução de um experimento científico com PAD, os conceitos relacionados a workflows servem como uma linguagem comum entre especialistas de domínio e cientistas da computação, habilitando pesquisas e esforços multidisciplinares. Entretanto, um conjunto de desafios ainda em aberto envolvem, por exemplo, a inclusão de especialistas de domínio nas iterações do ciclo de vida do experimento. Isso torna o *human-in-the-loop* (HIL) [W.F.J. 2007] um tópico bastante relevante, sendo atualmente abordado na comunidade acadêmica de workflows com dados em larga escala [Jagadish et al. 2014].

Trabalhos anteriores [Mattoso et al. 2015] abordam questões ligadas a este tópico sob o ponto de vista da proveniência de dados [Davidson and Freire 2008]. Neles, experimentos reais foram conduzidos com especialistas de domínio utilizando o Chiron, um motor de execução de workflows [Ogasawara et al. 2013]. Seus objetivos incluem: o uso efetivo de recursos de PAD; o suporte ao direcionamento dinâmico de workflows, para que workflows de longa duração possam ser ativamente gerenciados por especialistas de domínio; e consultas a uma base de proveniência de dados em tempo real.

Durante o ciclo de vida de um experimento científico, cientistas realizam a submissão de diferentes configurações de workflows até chegar a um resultado conclusivo. Para isso, analisam os resultados e frequentemente checam se fatos bem estabelecidos em seus domínios estão sendo respeitados. Caso o workflow em execução esteja refutando fatos conhecidos, ao invés de validá-los, é mais provável que ele seja interrompido, reconfigurado e só então recomeçado; a ciência normalmente incrementa o conhecimento previamente estabelecido ao invés de romper com os seus fundamentos. Durante esse processo, especialistas de domínio costumam recorrer à literatura, a anotações textuais e a resultados de experimentos anteriores. Assim, a checagem de fatos em workflows centrados em dados pode extrapolar a habilidade humana para administrar informações (i.e., relembrar fatos), sendo necessário interagir com fontes de dados externas e heterogêneas, de tabelas estruturadas a arquivos PDF.

De forma complementar, a construção de bases de conhecimento (CBC) aborda o conhecimento armazenado em fontes de dados altamente heterogêneas. Segundo Ré et al. [2014], "CBC é o processo de popular uma base de conhecimento (BC) com fatos (ou asserções) extraídos dos dados (e.g., texto, áudio, vídeo, tabelas, diagramas etc)". A ferramenta de código aberto DeepDive [Ré et al. 2014], para CBC, vem se destacando em uma diversidade de domínios tais como: paleontologia e geologia, genética e farmacogenética, no apoio a órgãos investigativos, e no enriquecimento da Wikipedia.

A motivação deste trabalho é aumentar o poder de análise de dados em tempo real de um Sistema de Gerência de Workflows Científicos (SGWfC), propondo a sua integração a um sistema automático de CBC. Esperamos, como resultado ainda não encontrado na literatura, prover uma solução declarativa para mineração, em tempo real, de fatos de domínio sobre dados de proveniência. Desta forma, durante a execução de um workflow em larga escala, podemos avaliar se os resultados obtidos até um dado momento respeitam os fatos conhecidos do domínio. Se necessário, o workflow pode ser reconduzido pelo cientista ou adaptado automaticamente. O objetivo é propor uma solução de integração de SGWfC e CBC que proporcione os seguintes benefícios: (i) poupar o tempo de cientistas com a inspeção manual *offline* dos dados de proveniência; (ii) economizar recursos de PAD, já que cientistas podem reconduzir um workflow à direção correta mais rapidamente; e (iii) ampliar o potencial de reprodutibilidade de um workflow, pois a publicação da BC usada naquele domínio complementaria a definição do próprio workflow e as consultas de proveniência, compondo um pacote abrangente para a reprodução do experimento.

Fora da literatura de workflows, destacamos que alguns métodos [Hu, Zhang et al. 2009-1, Hu, Sun et al. 2009-2] para clusterização de textos já se apóiam em bases de conhecimento externas para alcançar resultados superiores. Entretanto, essas utilizações são internas à clusterização. Em nossa proposta, a base de conhecimento é utilizada para validação externa dos resultados, como um especialista de domínio.

Este artigo está organizado da seguinte maneira: na seção 2 apresentamos os conceitos de CBC e o K-Chiron; na seção 3, apresentamos o caso de uso realizado e resultados experimentais; por fim, na seção 4, apresentamos as conclusões.

2. K-Chiron: integração de CBC a SGWfC

A solução de integração de um SGWfC e CBC é apresentada por meio do K-Chiron,

uma extensão do poder analítico de dados do Chiron para contemplar CBC. Três conceitos fundamentais definem a nossa abordagem para o K-Chiron, que são: (1) a estrutura básica de CBC; (2) uma breve organização das fontes de dados observadas ao longo do atual processo de depuração conduzido por especialistas de domínio; e (3) uma visão geral sobre como essas partes (CBC e a base de fatos) podem integrar um SGWfC.

2.1. Três camadas para CBC

De acordo com Ré et al. [2014], a perspectiva do usuário de um sistema de CBC pode ser minimamente definida por um sequência de três passos, como ilustrado na Figura 1. A figura ilustra um sistema de CBC destinado a popular uma BC sobre deputados federais que pertencem a bancadas suprapartidárias, com fatos extraídos de várias fontes de dados. Na figura, três deputados claramente associados à bancada empresarial, todos da unidade federativa ES, estão mencionados em uma entrada de dados não-estruturados (i.e., um texto), legível por humanos. O primeiro passo (passo 0) consiste no pré-processamento do texto: extração dos termos nas frases, classificação das Partes Do Discurso (POS) e reconhecimento de estruturas que tenham algum valor semântico útil ao domínio. O passo seguinte (passo 1) baseia-se no resultado anterior e corresponde à extração de características (*features*), incluindo: i. Extração de menções a deputados e bancadas (entidades para o sistema de CBC); ii. Candidatos a fatos verdadeiros (combinações de deputados e bancadas que possam estar associados); e iii. Características que ajudem a descrever tal relação, como as palavras entre as menções a essas entidades. Finalmente, o último passo (passo 2) demonstra uma regra de inferência unindo os passos anteriores. A regra é modelada segundo uma função lógica, e é o elemento de mais alto nível que um usuário de um sistema de CBC deve declarar antes que o sistema prossiga à fase de treinamento, processe os dados de entrada e gere fatos classificados como verdadeiros ou falsos, com seus respectivos níveis de certeza. Ao final desse exemplo, verificaríamos as bancadas com suas respectivas composições.

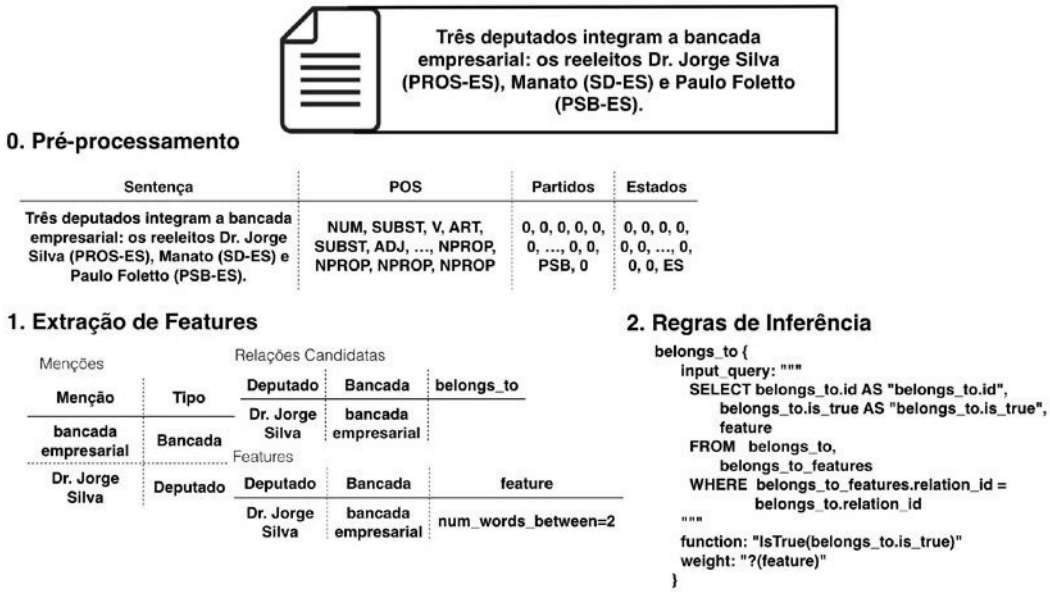

Figura 1. Adaptado de (Ré et al.): as três camadas para CBC.

2.2. A base de fatos: as fontes de dados do especialista de domínio

Como apresentado, especialistas de domínio mantêm seu conhecimento armazenado em

fontes de dados heterogêneas. Verificamos a existência de fontes de dados em três níveis de exploração do domínio: i. O mais interno diz respeito às anotações pessoais tomadas pelo próprio especialista; ii. O segundo nível representa fatos descobertos em conjunto com um grupo de pesquisa, frequentemente relacionados a explorações mais ambiciosas que exigem coordenação e comunicação entre membros do grupo, abarcando fatos não necessariamente publicados; iii. E o nível mais externo representa fatos conhecidos bem estabelecidos na comunidade científica daquele domínio. Assim, a base de fatos pode ser considerada como uma agregação de textos, de diversas tabelas de dados e de coleções de arquivos PDF, todos carregando fatos com diferentes níveis de confiança, tratando-se de um cenário recorrente à comunidade de CBC.

2.3. CBC no contexto do Chiron e os dados de proveniência

A arquitetura original do Chiron [Ogasawara et al. 2013] é apresentada na Figura 2 em conjunto com a proposta da base de fatos. O Chiron requer dados de entrada em um sistema de arquivos compartilhado (Figura 2, item 1), a definição de um workflow (item 2) e os usuários podem executar consultas sobre os dados de proveniência (item 3) armazenados em um banco de dados relacional durante a execução do workflow. O uso de um sistema de CBC adiciona à arquitetura do Chiron: uma base de fatos (item 5) e definições de características e de regras de inferência (item 4). As caixas azuis (itens 2 a 4) representam o que há de mais específico sobre o experimento e os itens numerados (1 a 5) representam o pacote completo do experimento, que poderia ser compartilhado para facilitar a reprodução do mesmo.

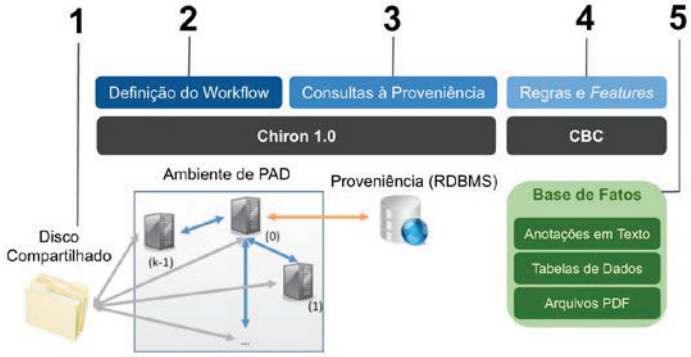

Figura 2. K-Chiron: CBC no contexto do Chiron e os elementos de dados.

A checagem de fatos por meio de dados de proveniência é uma atividade laboriosa e envolve interações com várias fontes de dados, uma vez que há geralmente muitos fatos importantes para um experimento. Em trabalhos anteriores [Dias et al. 2011, 2015], mostrou-se que a recondução de um workflow no Chiron pode, em tempo de execução, economizar um tempo substancial no ciclo de um experimento. A mineração de fatos deve ser pensada como um mecanismo para amplificar essa capacidade pois avalia automaticamente se os resultados obtidos respeitam os fatos conhecidos do domínio. Podem existir diferentes níveis de confiança para diferentes conjuntos de fatos conhecidos, cada nível causando decisões específicas na recondução do workflow.

3. Resultados Experimentais

O caso de uso realizado é uma aplicação no campo das Ciências Políticas, baseada em dados públicos disponíveis na Câmara dos Deputados brasileira. O seu propósito é processar dados brutos sobre como os 513 deputados federais votam em projetos de lei e

então permitir que um especialista de domínio identifique as bancadas mais proeminentes (i.e., que votam de maneira mais coesa) com o apoio de um workflow científico no Chiron. Como existem atualmente no Brasil 35 partidos políticos, o estudo das bancadas suprapartidárias torna-se especialmente interessante, embora desafiador por conta de sua natureza transversal à organização partidária oficial. Entende-se que o especialista de domínio queira interpretar agrupamentos de deputados federais que votam de maneira semelhante de acordo com as bancadas suprapartidárias mais tradicionais (agronegócio, empresarial, sindical, feminina e religiosa), ou a segmentos delas. Este experimento agrupa as votações na Câmara por meio de um algoritmo de clusterização [Hartigan & Wong 1979]. Como a quantidade de agrupamentos é desconhecida, o objetivo do workflow é encontrar a parametrização que resulta nos agrupamentos mais homogêneos em relação à identificação das bancadas.

O workflow é modelado para utilizar o algoritmo de clusterização K-Means [Hartigan & Wong 1979] da biblioteca Scikit Learn versão 0.17 [“Scikit Learn” 2016]. Variou-se o parâmetro *k* (i.e., a quantidade de agrupamentos) para varrer um espaço de busca pré-definido. Neste caso, com cinco bancadas de tamanhos variados, também convém testar se as bancadas não formariam subgrupos ou até mesmo um supergrupo diante das questões votadas. Assim, a varredura do parâmetro *k* foi definida no intervalo entre 4 e 18. Para fins de ilustração, a Figura 3 mostra uma visualização do K-Means executado com $k = 6$ e projetado para 2D (com uma análise dos componentes principais). A maneira escolhida para avaliar a qualidade dos agrupamentos nesta aplicação é pela medida de precisão [Manning et al. 2008]. Portanto, conduziremos a varredura em busca da maior precisão média, usando como *ground truth* a BC populada a partir do processo explicado no item 2.1.

O conjunto de dados usado neste experimento é extraído da API da Câmara [“Dados Abertos - Legislativo” 2016]. Ele cobre as votações de 384 deputados acerca de 14 projetos de lei considerados polêmicos tais como: reforma política, a diminuição da maioridade penal, o relaxamento de regulações do agronegócio, dentre outros. Os votos são computados como valores inteiros, conforme: sim (+2), obstrução (+1), não (0), abstenção (-1) e nulo (-2). A obstrução fica entre os valores "sim" e "não" porque é uma posição que, dependendo do contexto da votação, pode corresponder a um ou a outro. Também vale destacar que a lista com 384 nomes é um subconjunto dos 513 deputados eleitos, já que a parte que falta não é retornada pela API nas votações observadas.

3.1. O uso do DeepDive para CBC

Ao final de cada eleição, o DIAP (Departamento Intersindical de Assessoria Parlamentar) faz o lançamento de uma série de relatórios à imprensa analisando a composição das novas bancadas suprapartidárias. O departamento busca interpretar a nova Câmara dos Deputados independentemente das fronteiras partidárias, e tradicionalmente busca preencher os novos nomes das bancadas: a do agronegócio, a empresarial, a sindicalista, a feminina e a religiosa. Portanto, a BC deste exemplo consiste no entendimento dos seis relatórios lançados pelo DIAP e popular fatos do tipo "*deputado d_i pertence à bancada suprapartidária b_j*".

O uso do DeepDive começa com um passo de pré-processamento, que neste caso é feito sobre todos os relatórios do DIAP. Em conformidade com o exemplo oficial do DeepDive, usamos a biblioteca NLTK versão 3.1 [“Natural Language Toolkit” 2016] para

realizar os passos básicos para o processamento de linguagem natural: a extração de todas as frases dos relatórios com seus respectivos termos; a classificação de cada POS usando o corpus MacMorpho do NLTK, que consiste em um conjunto anotado de textos de notícias (mais de um milhão de termos) extraídas de diversos jornais brasileiros. Além disso, reconhecemos dentro das frases dois conjuntos de termos: as abreviações de unidades federativas brasileiras e os acrônimos dos partidos políticos, pois ambas informações são úteis durante a detecção de entidades. Portanto, para cada frase pré-processada deve existir: a frase bruta; os termos dela extraídos; os termos seguidos de suas classificações como POS; os termos identificados como unidades federativas brasileiras; e os termos identificados como acrônimos de partidos políticos.

Uma vez que o passo de pré-processamento está concluído, construímos dois extratores, criando código especificamente para esta BC. O primeiro é o extrator de deputados, que aplica algumas heurísticas para extrair candidatos a menções – assume que, em relatórios, deputados são mencionados próximos de seus respectivos partidos políticos e de suas unidades federativas originárias. Para ampliar os candidatos a menções, o extrator considera todos os unigramas, bigramas e trigramas anteriores à heurística apresentada (i.e., um acrônimo de partido político seguido de um estado brasileiro, ou vice-versa). Já o segundo extrator é o que obtém as bancadas suprapartidárias. Similarmente, ele aplicar heurísticas para extrair candidatos a menções – assume que essas entidades são caracterizadas pelo termo "*bancada*" seguido de um termo classificado como adjetivo. Os extratores foram implementados em Python e declarados na configuração do DeepDive.

Com todos os candidatos a menções a entidades extraídos, seguimos para extrair candidatos à relação *belongs_to*. Para o propósito do workflow a ser executado após a CBC, todas as co-ocorrências de deputados e bancadas em uma mesma seção são consideradas candidatas à relação-alvo (*belongs_to*). Entretanto, este extrator aplica um filtro adicional: ele remove candidatos à relação-alvo se o deputado em questão não constar entre os 384 nomes obtidos pela API da Câmara, admitindo uma distância de Levenshtein de até 2 [Navarro 2001]. Assim, todos os fatos que serão obtidos a partir dos candidatos à relação estarão ancorados aos dados de entrada do workflow; só conheceremos as bancadas de quem interessar ao experimento.

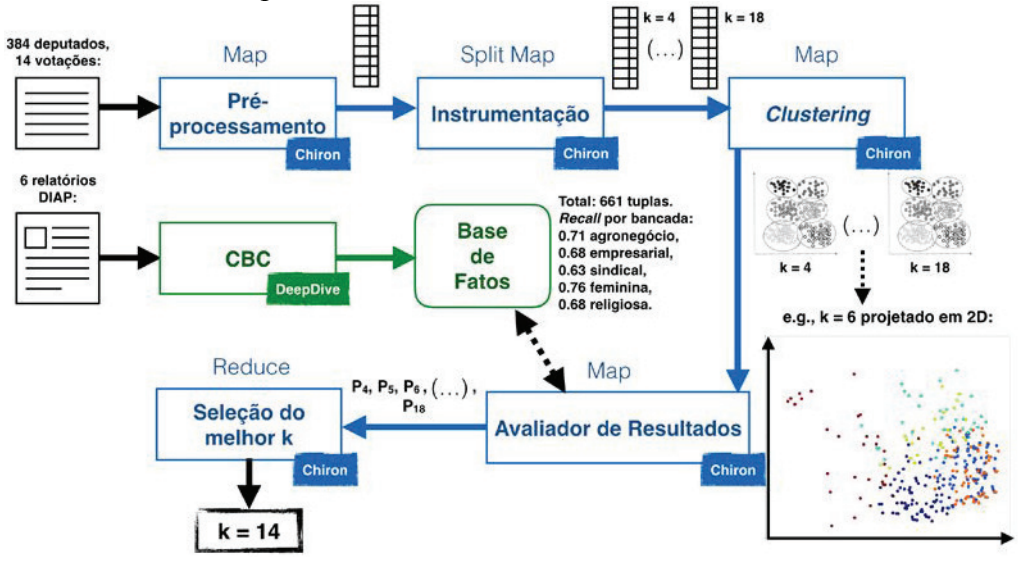

Figura 3. O workflow analítico da Câmara com o K-Chiron.

Para treinar o modelo estatístico do DeepDive, compila-se uma lista com 60 fatos verdadeiros (i.e., deputados específicos que de fato pertencem a bancadas específicas). A geração de uma lista de fatos verdadeiros constitui a chamada "supervisão distante", necessária ao DeepDive. Além disso, cria-se um extrator de características genéricas para cada sentença extraída como uma candidata à relação *belongs_to*. Tal extrator emprega a biblioteca Python fornecida pelo DeepDive, chamada DDLib, para emitir características genéricas a partir do texto das frases. O uso da DDLib para extrair features é a abordagem recomendada para tratar de atributos de texto no DeepDive. Por fim, declara-se uma regra de inferência no arquivo de configuração do DeepDive conforme apresentada na Figura 1. Ela ao mesmo tempo captura os exemplos de fatos verdadeiros e, nos casos preditos pelo DeepDive, retorna como saída o valor da predição. A regra de inferência popula a BC tanto com os fatos conhecidos usados no treinamento, quanto com os fatos preditos como verdadeiros ou falsos a partir dos candidatos extraídos à relação *belongs_to* com uma cobertura de 63% a 76% das bancadas desejadas.

3.2. O workflow integrado à base de conhecimento

O workflow do experimento com o K-Chiron são descritos na Figura 3. Nesta prova de conceito, a integração com o DeepDive foi feita a partir de uma atividade do workflow programada para realizar consultas à base de fatos no DeepDive. Em um cenário mais geral, as atividades do workflow no K-Chiron podem ser configuradas de maneira declarativa para minerar os dados de proveniência, checando os fatos na BC. Neste caso, o SGWfC é responsável pela varredura do espaço de busca enquanto a base de fatos possibilita a verificação dos resultados em tempo real. Resumindo a sequência de passos para o workflow, temos: i. Uma atividade do tipo *Map* consolida as votações; ii. Uma atividade do tipo *Split Map* instrumenta o conjunto de dados para varrer todo o espaço de busca de *k*; iii. Uma atividade *Map* lê as votações e aplica o K-Means para cada valor distinto de *k*; iv. Cada resultado da clusterização, em outra atividade *Map*, é testado com respeito à BC populada pelo DeepDive, culminando em uma medida de precisão para cada agrupamento e uma precisão média para cada valor de *k*. Por fim, o workflow termina indicando para qual valor de *k* a precisão média é ótima, significando a produção dos agrupamentos mais homogêneos (em que proporção, na média, os agrupamentos foram preenchidos por membros de uma única bancada suprapartidária).

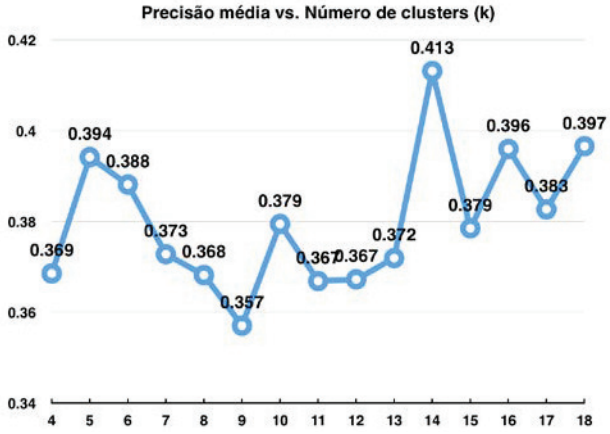

Figura 4. A precisão média dos agrupamentos variando k entre [4,18].

A Figura 4 apresenta a qualidade média dos agrupamentos e pode-se observar que a precisão ótima é alcançada com $k = 14$, lembrando que tal verificação diz respeito à

base de fatos obtida com o DeepDive. Se o especialista de domínio fosse navegar pelos agrupamentos resultantes para todos os valores de *k* testados, sua avaliação passaria pelos 384 nomes de deputados 15 vezes (dado que |[4,18]| = *15*). Isso significaria uma avaliação de quase seis mil nomes, possivelmente consultando os seis relatórios do DI-AP nos quais são armazenados um total de 582 fatos relevantes (i.e., deputados específicos pertencem a bancadas suprapartidárias específicas).

4. Conclusões

Encapsular o conhecimento do especialista de domínio usando CBC, e com isso levá-lo para dentro do ciclo do workflow, mostrou ser uma ideia bastante produtiva. É difícil mensurar com precisão a quantidade de tempo efetivamente poupada, mas é importante destacar que, de seis mil avaliações originalmente demandadas, apenas uma única avaliação precisou ser feita tirando proveito das consultas em tempo real. Aplicar o DeepDive ao HIL de workflows pode ser uma ideia ainda mais produtiva se explorada em cenários de execução dinâmica de workflows, em problemas que também dependam da administração de uma grande quantidade de fatos por parte do especialista de domínio. Nesses casos, a mineração de fatos em larga escala pode: i. poupar o tempo de cientistas com a inspeção manual dos dados de proveniência; e ii. economizar recursos de PAD, já que cientistas poderiam reconduzir um workflow à direção correta mais rapidamente.

5. Referências

Dados Abertos - Legislativo (2016). http://www2.camara.leg.br/transparencia/dados-abertos/dados-abertos-legislativo/dados-abertos-legislativo, [accessed on Apr 7].

Davidson, S. B. and Freire, J. (2008). Provenance and Scientific Workflows: Challenges and Opportunities. In *Proceedings of the 2008 ACM SIGMOD*

Dias, J., Guerra, G., Rochinha, F., et al. (may 2015). Data-centric iteration in dynamic workflows. *Future Generation Computer Systems*, v. 46, p. 114–126.

Dias, J., Ogasawara, E., Oliveira, D., et al. (2011). Supporting Dynamic Parameter Sweep in Adaptive and User-Steered Workflow. In WORKS '11. ACM.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), p. 100–108.

Hu, X., Sun, N., Zhang, C., & Chua, T. S. (nov 2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 919-928).

Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (june 2009). Exploiting Wikipedia as external knowledge for document clustering. In Proceedings of the 15th ACM SIGKDD (pp. 389-396).

Jagadish, H. V., Gehrke, J., Labrinidis, A., et al. (1 jul 2014). Big data and its technical challenges. *Communications of the ACM*, v. 57, n. 7, p. 86–94.

Manning, C. D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mattoso, M., Dias, J., Ocaña, K. A. C. S., et al. (may 2015). Dynamic steering of HPC scientific workflows: A survey. *Future Generation Computer Systems*, v. 46, p. 100–113.

Natural Language Toolkit (2016). http://www.nltk.org/, [accessed on Apr 7].

Navarro, G. (mar 2001). A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, v. 33, n. 1, p. p 31–88.

Ogasawara, E., Dias, J., Silva, V., et al. (2013). Chiron: A Parallel Engine for Algebraic Scientific Workflows. *Concurrency and Computation*, v. 25, n. 16, p. 2327–2341.

Ré, C., Sadeghian, A. A., Shan, Z., et al. (23 jul 2014). Feature Engineering for Knowledge Base Construction. *arXiv:1407.6439 [cs]*,

Scikit Learn (2016). http://scikit-learn.org/stable/, [accessed on Apr 7].

W.F.J., B. (2007). Human-in-the-loop'simulation: the right tool for port design. In *Port Technology International*.