UNIVERSIDADE FEDERAL DE SANTA CATARINA

tCALC: Agrupamento de Currículos Lattes por Afinidade de Áreas de Conhecimento Considerando Temporalidade

Jaime Mendes da Silva

Florianópolis,
2016.

UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

tCALC: Agrupamento de Currículos Lattes por Afinidade de Áreas de Conhecimento Considerando Temporalidade

Jaime Mendes da Silva

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Florianópolis,
2016.

Jaime Mendes da Silva

tCALC: Agrupamento de Currículos Lattes por Afinidade de Áreas de Conhecimento Considerando Temporalidade

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

_____

Rodrigo Gonçalves
Orientador

_____

Carina Friedrich Dorneles
Co-orientadora

Banca Examinadora:

_____

Roberto Willrich

_____

Ronaldo dos Santos Mello

SUMÁRIO

LISTA DE FIGURAS

LISTA DE REDUÇÕES

UFSC Universidade Federal de Santa Catarina

CNPq Conselho Nacional de Desenvolvimente Científico e Tecnológico

CT&I Ciência, Tecnologia e Inovação

HTML *HyperText Markup Language*

XML *Extensible Markup Language*

CONSCIENTIAS Comunidade para Ontologias em Ciência, Tecnologia e Informações do Aperfeiçoamento de Nível Superior

LMPL Linguagem de Marcação da Plataforma Lattes

MAP *Mean Average Precision*

KDD *Knowledge Discovery in Databases*

LExR *Lattes Expertise Retrieval*

API *Application Programming Interface*

RESUMO

Trabalhos realizados anteriormente, através de algoritmos de *clustering* de dados, agruparam currículos Lattes de profissionais da área de ciência, tecnologia e inovação [1]. Os grupos gerados por esse processo evidenciavam informações sobre a área de atuação desses profissionais e quais pertencem a uma mesma área. O presente trabalho estende o que foi realizado ao analisar o impacto de qualidade e performance causado pela consideração do fator tempo no processo de agrupamento dos currículos. A inclusão da temporalidade vem da evidência na literatura de que aplicações de busca por competências se beneficiaram da mesma [4]. A aplicação dá-se pelo fato de que profissionais que atuaram em determinada área do conhecimento no passado podem não ser mais atuantes na mesma.

Palavras-chave: Agrupamento de dados. *Knowledge Discovery in Databases*. Engenharia do Conhecimento. Plataforma Lattes. Temporalidade.

1. INTRODUÇÃO

Dados são elementos chave de sistemas computacionais. Podemos colocar de maneira enfática que os computadores atuais existem, em sua instância mais básica, para processar dados, sejam esses originados em aplicações científicas ou domésticas, gerados na web ou por equipamentos sofisticados de medição, utilizados para auxiliar em tarefas do dia-a-dia ou para questões profissionais com enorme impacto financeiro [8]. A importância dos computadores nas atividades humanas, somada ao grande avanço tecnológico dos recursos computacionais e sua consequente redução de custos, trouxe como consequência um estado de geração rápida, variada e massiva de dados [8].

Produzir dados não garante por si só que informação seja adquirida a partir deles e que algum conhecimento seja obtido a partir dessa informação. Para preencher essa lacuna, surgiram diversas disciplinas que se propõem a tratar os dados de forma a torná-los úteis para o uso humano em suas diversas aplicações. Entre elas, a Análise de Dados e suas diversas técnicas [13].

Uma importante técnica para o escopo deste trabalho é a Mineração de Dados (do inglês *data mining*). Ela consiste do processo de descoberta de padrões interessantes, tal como da criação de modelos preditivos e descritivos acerca de uma grande quantidade de dados [2]. Esses padrões e modelos são alcançados através da execução de atividades bem definidas que fazem com que esses dados revelem propriedades invisíveis inicialmente. Uma dessas atividades é o agrupamento (*clustering*), que procura particionar os dados evidenciando os chamados grupos naturais, ou seja, os grupos que concentram todas aquelas amostras que demonstram comportamentos similares em relação a determinadas propriedades [2]. O agrupamento é um dos temas centrais deste trabalho.

Como a principal motivação da mineração de dados é obter um conhecimento específico a partir de dados, é natural que ela seja aproveitada para diversos tipos de aplicações, utilizando-se de diversos tipos de dados. Um desses tipos, os dados textuais encaixam-se na categoria de dados não-estruturados e são frequentemente alvos de mineração por: (1) serem bastante comuns nos processamentos feitos hoje com computadores e (2) por normalmente possuírem grande potencial semântico para aquisição de conhecimento [3].

A fonte de dados utilizada neste trabalho são currículos da plataforma Lattes, que reúne bases de dados de currículos profissionais da área de pesquisa, desenvolvimento e inovação no Brasil. Ela se encaixa na categoria de dados semi-estruturados, onde seu conteúdo se manifesta através de textos. Dessa forma, no processo de descoberta de conhecimento conduzido, essa estrutura será considerada apenas na etapa inicial de transformação. Depois disso, os dados serão tratados como texto. A informação que será buscada a partir deles são as competências dos autores desses currículos.

Este trabalho parte da obra "CALC: Agrupamento de Perfil Científico por Afinidade de Áreas de Conhecimento Utilizando Currículos Lattes", por Bernardo de Farias Esteves (2015). Enquanto a preocupação da obra citada era a de implementar o agrupamento de currículos Lattes por afinidade nas suas áreas de conhecimento, utilizando-se do *framework* URSA, este trabalho busca ir além adicionando o caráter da temporalidade no agrupamento dos currículos apesar de manter o fator competências dentro do cálculo de similaridade.

A introdução do aspecto temporal ao problema que já havia sido abordado no projeto CALC dá-se com base em linhas recentes de pesquisa que propõem que o tempo é um fator com alto teor de relevância para avaliar a perícia da qual um profissional dispõe em relação a uma área [4].

As diferenças, motivos e relevância das implementações são elucidados no restante da obra e, à medida que forem apresentados os tópicos abordados por este trabalho que tiveram como base o CALC, é apresentada a implementação anterior destacando o que foi adicionado ou modificado.

1.1. OBJETIVOS

1.1.1. Objetivo Geral

Ao fim deste trabalho espera-se contribuir para o conhecimento sobre os efeitos da temporalidade na obtenção de padrões em dados de texto minerados. Busca-se conseguir isso através da implementação de uma versão alternativa (que considere o fator tempo) de uma aplicação que elabora o agrupamento de dados de currículos da plataforma Lattes. Inclui-se neste escopo a produção de uma revisão da literatura sobre o campo de estudos, além da experimentação e análise sobre o impacto da adição do aspecto temporal na aplicação mencionada..

1.1.2. Objetivos Específicos

Os objetivos específicos são os seguintes:

1. A partir do algoritmo CALC, implementar uma variação temporal – tCALC;
2. Gerar agrupamentos feitos por ambos CALC e tCALC, expostos a diversos cenários de teste;
3. Integrar a aplicação com a base de dados LExR para qualificar o agrupamento através de dados adicionais sobre os currículos Lattes;
4. Analisar os resultados e a qualidade do algoritmo em relação aos agrupamentos gerados.

1.2. METODOLOGIA

Para que se alcance os objetivos deste trabalho, os seguintes métodos são contemplados no projeto:

- Fundamentação teórica que proporcione a delimitação e definição precisa dos conhecimentos utilizados para desenvolvimento do trabalho;
- Formalização do projeto da aplicação que permita a garantia de conformidade do procedimento adotado para com o descrito na literatura;
- Implementação da aplicação;
- Experimentação do aplicativo a partir de dados reais de entrada;
- Análise quantitativa dos resultados dos experimentos conforme métricas previamente estipuladas.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são abordados os principais temas sobre o qual o projeto do tCALC apoia-se e aos quais ele busca contribuir. Por mais que esta obra tenha se preocupado em um aspecto prático de mineração e um aspecto analítico sobre a área de competências, todas as atividades foram executadas com a preocupação em buscar respaldo na literatura científica.

Pretende-se que, ao final da fundamentação teórica, o leitor tenha adquirido familiaridade suficiente com os conceitos apresentados para que a relevância das discussões apresentadas na descrição do processo de mineração fique evidente.

2.1. CURRÍCULOS LATTES

2.1.1. A Plataforma Lattes

A plataforma de currículos Lattes é uma proposta do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) para unificação e integração dos currículos de todos os tipos de profissionais com envolvimento no setor de pesquisas tecnológicas e científicas. Essa plataforma é um sistema de informação que permite consultar currículos (chamados “currículos Lattes”) dos profissionais tal como a gestão, por eles, de seus próprios currículos, armazenados em um banco de dados mantido pelo CNPq [14].

O caráter integrador do Lattes e o alcance do CNPq contribuíram para que a adoção fosse ampla dentre as instituições de fomento à pesquisa, tornando a plataforma um padrão nacional. Ela permite que quaisquer interessados em analisar os perfis e realizações de pesquisadores e estudantes de nível superior, tais como investidores e responsáveis por pequisa em geral, possam alcançar facilmente essas informações [14].

Além dos currículos, a plataforma Lattes possui bases de dados sobre grupos de pesquisa em atividade no Brasil e de instituições que mantêm relações com o CNPq chamados, respectivamente, de Diretório dos Grupos de Pesquisa e Diretório de Instituições. Tais grupos podem ser conferidos no site oficial da plataforma (www.lattes.cnpq.br). Este trabalho, por não se utilizar desses diretórios e seus dados, não trará aprofundamento acerca deles. O interesse maior para esta obra são os currículos.

O padrão de currículos Lattes foi concebido para fazer contraponto aos tradicionais *curriculum vitae* no que diz respeito ao detalhamento das informações sobre experiências e realizações acadêmicas e profissionais de um indivíduo. Enquanto o *curriculum vitae* costuma ser conciso, o currículo Lattes busca uma descrição completa e precisa da extensão dos estudos e atribuições exercidas pelo profissional. O interesse por esse requisito de detalhamento faz-se presente a partir da necessidade, por parte dos financiadores da pesquisa, de um conhecimento preciso sobre as capacidades dos pesquisadores [14].

2.1.2. Conteúdo de um Currículo Lattes

A partir da observação dos campos que a plataforma proporciona no ato de criação e edição de um currículo, pode-se identificar as informações disponíveis em um currículo Lattes. A partir dessa análise, é possível a descrição dos principais campos disponíveis abordados nesta obra à medida que se discute a implementação, visto que o Lattes proporcionará os dados a serem utilizados para definição dos perfis de competências.

Por ordem de apresentação, os seguintes campos estão disponíveis em um currículo Lattes:

1. **De caráter pessoal:**
    a. **Resumo:** descrição breve das principais funções e realizações, normalmente situando-as na época e locais nos quais se passaram.
    b. **Identificação**: apresenta o nome completo e os possíveis nomes utilizados em citações bibliográficas.
    c. **Endereço**: local onde pode ser encontrado e informações gerais para contato.
2. **De caráter profissional e acadêmico:**
    a. **Formação acadêmica/titulação**: informações sobre a formação como nível, curso, instituição, trabalho que consagrou a formação (trabalho de conclusão de curso, teses, dissertações), áreas (para pós-graduação), orientadores e palavras-chave mais relevantes das obras apresentadas para titulação e o período de envolvimento com o curso.
    b. **Atuação profissional**: traz as instituições nas quais o profissional atuou, detalhando o vínculo existente em um dado período e especificando as atividades exercidas.
    c. **Linhas de pesquisa**: enumera as linhas de pesquisa do autor e estabelece objetivos desse para com a linha, além de palavras-chave dos estudos exercidos.
    d. **Projetos de pesquisa**: lista os projetos de pesquisa dos quais o pesquisador fez ou faz parte, expressando o período de envolvimento, uma descrição detalhada do projeto, situação atual do andamento do

projeto, natureza, quantidade e nível acadêmico dos envolvidos, financiadores e quantidade de produções geradas através do projeto.

e. **Projetos de extensão**: enumera, descreve e localiza temporalmente os projetos de extensão realizados.
f. **Prêmios e títulos**: lista os prêmios do autor, indicando o ano e evento em que se passaram.
g. **Produções**: lista completa de publicações tais como artigos publicados em periódicos, livros, capítulos de livros, textos em jornais ou revistas, trabalhos e resumos publicados em anais de congressos, apresentações, teses, entre outros tipos de produções. Para cada publicação são listados autores, título, ano, localidade e eventos. Além de obras publicadas, também são apresentadas obras de caráter técnico que foram produzidas pelo autor, tais como programas de computador, projetos de engenharia, assessorias, consultorias, entre outras.
h. **Bancas**: lista de bancas das quais o pesquisador foi parte, identificando o nível acadêmico do trabalho, os demais membros e o nome da obra.
i. **Eventos**: enumeração de eventos voltados à ciência, tecnologia e inovação dos quais o indivíduo participou ou organizou, tais como congressos, exposições ou feiras.
j. **Orientações**: apresenta todos os trabalhos de diversos níveis acadêmicos os quais foram orientados pelo autor do currículo, trazendo informações da obra tais como autor, título, instituição e ano.

Essa extensa lista das partes integrantes de um currículo Lattes reafirma o detalhamento e a relevância desse modelo de currículos quanto à descrição da carreira do profissional de pesquisa retratado. Essa relevância faz da plataforma Lattes uma fonte interessante de informações para estudos sobre o perfil dos profissionais que atuam na área de ciência, tecnologia e inovação (CT&I) no Brasil.

2.1.3. Estrutura do Currículo Lattes

Um currículo Lattes pode ser encontrado disponível na plataforma nos formatos HTML e XML, esse último idealizado e proposto a fim de promover "uma forma comum de troca de informações entre agências de fomento e suas instituições usuárias", como é descrito na página da Comunidade para Ontologias em Ciência, Tecnologia e Informações do Aperfeiçoamento de Nível Superior (CONSCIENTIAS)[1], grupo dedicado ao desenvolvimento de ontologias que contribuam para a troca de informações entre instituições interessadas em CT&I e que integra a LMPL (Linguagem de Marcação da Plataforma Lattes), equipe que é responsável pelo desenvolvimento das gramáticas XML utilizadas na implementação da Plataforma Lattes.

A finalidade da disponibilização dos currículos em formato XML descrita acima reflete o potencial para servir de material a aplicações diversas que possuam relevância a financiadores, instituições da área de CT&I e, de maneira geral, para todo o mercado, nos seus ramos que podem se beneficiar de informações sobre especialistas. Este trabalho, inclusive, faz uso desse material em formato XML através de método e por motivação trazidos em outras seções desta obra.

A estrutura de um currículo Lattes no formato XML respeita aproximadamente os campos descritos na seção anterior, organizando-os em uma árvore que submete as seções mais específicas às mais gerais. A estrutura típica de um arquivo XML é apresentada mais adiante neste trabalho, incluindo exemplos que reiteram e reforçam o que foi introduzido nesta seção.

[1] http://lmpl.cnpq.br/lmpl/

2.2. COMPETÊNCIAS

Entre as definições trazidas nos dicionários para a palavra competência podemos encontrar: “capacidade que um indivíduo possui de expressar um juízo de valor sobre algo a respeito de que é versado; idoneidade; soma de conhecimentos ou habilidades”. Nessa definição se apoia a ideia de competências abordada por esta obra. Para expressar de forma mais condizente, derivando-se dessas definições, temos competências como o conjunto de conhecimentos dos quais um profissional dispõe ou áreas do conhecimento nas quais ele tem aptidão para atuar. Essa última parte é colocada com a finalidade de gerar uma abstração necessária para a classificação de conhecimentos distintos, porém próximos, sob uma mesma visão.

Para clarificar essa definição, podemos nos utilizar de um exemplo. Dados dois profissionais especializados em Computação, um dos quais dedica sua carreira ao estudo de bancos de dados relacionais e o outro se dedica aos não-relacionais, podemos, dependendo da nossa finalidade, afirmar que eles podem contribuir para um estudo sobre bancos de dados. Portanto, através desse exemplo, fica evidente que pode haver a necessidade de se tratar conhecimentos distintos sob uma mesma área de conhecimento. Já a especificidade (ou generalidade em contraponto) que podemos exigir de uma classificação dessas pode variar muito. A partir disso, esta obra discute a tomada de diferentes medidas de similaridade, pois essas medidas causam impacto na especificidade ou generalidade da classificação de um conhecimento em uma área.

2.2.1. Relevância

Uma discussão sobre profissionais e suas especialidades é praticamente indissociável de uma discussão sobre o mercado e suas demandas acerca das competências de profissionais. É nesse âmbito que este trabalho, tal como aquele no qual se baseia, tenta adentrar. Tendo em vista as principais aplicações do agrupamento de dados e os motivos associados a sua concepção, podemos facilmente detectar sua aplicabilidade junto ao agrupamento de profissionais de

mesmas competências, permitindo formar grupos de trabalho sobre determinadas áreas de conhecimento. O uso pode estender-se ainda para profissionais que queiram desenvolver uma rede de contatos com especialistas da mesma área [1].

Portanto, para satisfazer essas e quaisquer outras aplicações que possam haver em decorrência do agrupamento por competências, precisamos ser capazes de elaborar o agrupamento. Essa capacidade já foi provida pela aplicação CALC e foi preservada na medida que o tCALC se desenvolveu a partir dela, sempre mantendo no foco as competências como características mais importantes do agrupamento, porém incluindo o tempo como uma variável.

2.2.2. Como identificar competências em um currículo

Como foi definido na seção anterior, o conceito de competência pode ser visto em diferentes níveis de abstração. Portanto, não é trivial, dado um pesquisador, enumerar quais as suas competências. Por exemplo, um especialista em Direito Tributário é também um profissional de Direito. Portanto, podemos escolher ser específicos e dizer que um determinado jurista possui competência em Direito Tributário ou podemos ser mais genéricos ao dizer que a mesma pessoa possui competência em Direito.

No caso dos currículos Lattes, como foi descrito anteriormente, existem diversos campos a serem preenchidos pelo autor de um currículo. Muitos deles podem ser utilizados para se obter informações que podem ser aproveitadas para estimar competências. Alguns podem conter informações que remetam a competências mais genéricas, outros podem trazer informações mais específicas ou até mesmo trazer vários níveis de abstração. O campo Áreas de Atuação, por exemplo, pode especificar até quatro níveis de abstração para uma dada área de atuação, denominados: grande área, área, subárea, especialidade.

Outros campos dos currículos podem não trazer explicitamente uma competência, mas possibilitam que, através de técnicas de análise de dados, possamos extrair informações que nos ajudem a identificá-la. Isso vale para campos contendo título de publicações, bancas ou eventos, por exemplo.

Nesta obra, visando obter informações sobre competências e sempre mantendo uma relação com um período no tempo onde tal área de atuação se fez presente na carreira de um profissional, foi escolhida a abordagem de analisar os títulos de publicações presentes no campo Produções do currículo Lattes.

A abordagem inicial, entretanto, visava utilizar, a exemplo do CALC, os campos diretamente relativos a áreas do conhecimento. Porém, percebeu-se que essas informações não tinham a informação temporal necessária para a nova aplicação. A partir daí optou-se pela seção Produções mencionada anteriormente.

Em outros trechos da obra essa escolha é discutida em mais detalhes, mas o importante neste momento é se fazer notar esse caráter implícito do campo Produções – mais especificamente do título de uma produção – quanto à disponibilização da competência de um autor. Um autor que tenha publicado sobre um determinado tema certamente tem autoridade naquela área.

2.3. TEMPORALIDADE

A busca por competências, assim como diversas outras aplicações da mesma natureza, ao direcionar sua análise nas variáveis mais estritamente relacionadas com a finalidade da busca, acaba ignorando fatores importantes como a informação temporal. Normalmente isso pode se confirmar como um equívoco, afinal o fator tempo pode revelar uma grande quantidade de informações relevantes dependendo do conhecimento que se procura obter a partir da busca, o que pode agregar qualidade à atividade [4].

2.3.1. Aplicações da Temporalidade na Recuperação de Informação

Muitas aplicações do mundo real podem exigir ou se beneficiar de uma análise considerando o tempo. Em alguns casos clássicos como na mineração de dados em bancos de dados temporais essa característica fica evidente, mas em alguns casos as vantagens de se considerar o atributo tempo passam despercebidas.

Um exemplo que não se aproxima tanto do caso que é tratado neste trabalho, mas que pode trazer uma visão da importância do tempo em certas aplicações que envolvem o tratamento de informações, é o de um indexador de notícias. Em uma ferramenta que busca notícias do banco de dados de um jornal que não considera a data das notícias, diante da pesquisa sobre o termo “eleições presidenciais no Brasil” poderia, através de seus critérios de indexação, exibir notícias muito antigas referentes ao tema, talvez notícias sobre as eleições de 1989. Esse provavelmente seria um resultado irrelevante para a maioria dos usuários. Já se a ferramenta considerasse o fator tempo e exibisse as notícias mais recentes sobre o tema, é bem mais provável que os resultados fossem satisfatórios para a maioria dos leitores.

Já alguns autores propõem que usuários de buscadores podem se beneficiar de outras abordagens diferentes da exposta acima. Alonso e Gertz (2006), por exemplo, em relação ainda sobre buscadores, argumentam que apenas ordenar as notícias pela ocorrência pode não ser suficiente para oferecer um uso confortável ao leitor. Eles propõem o emprego de *clustering* em relação a buscadores para que

esses ofereçam grupos relevantes e coesos de resultados em uma mesma pesquisa. No exemplo que utilizam em sua obra eles trazem a busca pelo termo “copa do mundo”. Se o buscador de notícias apenas ordenar os resultados exibindo os mais recentes primeiro, é provável que resultados sobre a última copa do mundo ou futuras copas populem a saída, considerando-se critérios usuais. Tendo isso em vista, alguns usuários que procuram por notícias sobre outras copas não teriam fácil acesso a elas através da busca por esses termos. Se o buscador utilizasse técnicas de agrupamento, ele poderia gerar *clusters* de saída, cada um referente a uma determinada copa do mundo e possibilitando que o usuário possa escolher sobre qual delas ele quer ler.

Essa última aplicação assemelha-se mais com o uso da temporalidade nesta obra, visto que ambos utilizam *clusters* baseados no tempo, embora os fins sejam diferentes. Além das abordagens citadas nesta seção, a literatura apresenta outras que podem se fazer relevantes principalmente no estudo sobre buscadores [15]. Esta obra, porém, limita-se em citá-las apenas para evidenciar a aplicabilidade e a variedade de ópticas sob as quais pode ser visto o uso da temporalidade.

2.3.2. Manifestação do Tempo em Dados Textuais

Uma característica do fator tempo que se mostra relevante para o debate é como ele se manifesta em relação a dados como as notícias usadas como exemplo anteriormente e os currículos trabalhados neste projeto. É importante que se faça lembrar que os dados textuais são usualmente não-estruturados ou no máximo semi-estruturados. Tendo isso em vista, Alonso e Gertz (2006) atestam que informações temporais se apresentam, principalmente, de 3 maneiras em um texto: (1) explicitamente, através de unidades de tempo como datas (“21 de Junho de 2001”, “Maio de 2014”, “1982”); (2) implicitamente, através de palavras que indiquem uma data específica (“natal de 1988”, “dia da abertura das olimpíadas de 2016”); (3) vagamente, através de dados que por si só não revelam a informação temporal e necessitam acompanhar manifestações explícitas ou implícitas para isso (“na última quinta-feira”, “ontem”).

Além dessa característica sintática dos atributos temporais em textos, pode-se também analisar um outro viés que diz respeito ao conteúdo semântico inerente aos atributos de temporalidade que cada um naturalmente produz. Dependendo da natureza do texto, podemos inferir ou ao menos esperar que os dados sobre tempo pertençam a um domínio bem definido. Textos sobre atividades financeiras, por exemplo, comumente descrevem eventos que ocorrerão em um futuro próximo. Currículos, por outro lado, costumam descrever evento passados e estabelecer uma linha do tempo [5].

No que diz respeito ao conjunto de dados usados nesta obra, todos os dados temporais identificados apresentam-se de maneira explícita e normalmente em granularidade anual, o que é suficiente para esta aplicação. Como citado anteriormente, busca-se nesses dados uma sequência bem definida de eventos que se passaram na carreira dos profissionais, possibilitando que se analise as variações ocorridas (e.g. as mudanças de área de atuação ou competência).

2.3.3. Estratégias de Uso do Tempo na Recuperação de Informação

A utilização de temporalidade na recuperação de informação dá-se de diversas formas. Alonso e Gertz (2006) utilizam-se disso para gerar grupos de notícias produzidas naquele ano e indexadas pelo termo consultado. Li e Tang (2008) utilizam junto a uma modelagem estatística de *random walk* (passeio aleatório). Esses são os métodos utilizados como base teórica deste trabalho que apresenta sua própria abordagem. Pode ainda haver outras abordagens na literatura que podem ser consideradas para o problema de recuperação de informação, mas essas não são consideradas nesta obra.

Neste projeto a estratégia adotada se dá através da partição do conjunto de termos utilizados para testar similaridade de acordo com o ano das publicações de um autor que carregam aqueles termos. Após isso, realiza-se o agrupamento para cada ano, biênio ou quaisquer medidas abordadas e discutidas neste trabalho.

2.3.4. Experimentos e Resultados da Literatura

Dos dois trabalhos referidos na seção anterior, apenas o segundo, de Li e Tang (2008) conduziu experimentos com os modelos gerados. O primeiro trabalho, de Alenso e Gertz (2006) apresentou um protótipo de indexador-agrupador e deixou as análises qualitativas em aberto.

Dos experimentos elaborados pelos autores, foram avaliadas métricas como R-precisão, revocação e MAP (*Mean Average Precision*). A primeira métrica avalia qual a taxa de um número arbitrário (R) de elementos considerados relevantes pelo algoritmo em relação a todo o conjunto de dados. A segunda avalia a taxa dos documentos considerados relevantes em relação à quantidade total de documentos relevantes do conjunto de dados. A última métrica avalia adicionalmente a ordem dos documentos recuperados [6].

Não cabe a esse trabalho explicitar os resultados adquiridos pelos autores mas podemos fazer uso das conclusões deles. A avaliação que fizeram dos resultados obtidos foi a seguinte: “resultados experimentais demonstram que melhorias podem ser obtidas em comparação aos métodos tradicionais”. Essa evidência apoia a suspeita de que o emprego do fator tempo na aplicação com a qual esta obra se preocupa pode ser tão benéfico quanto foi em relação ao trabalho de referência.

2.3.5. Temporalidade *versus* Competências

No trabalho publicado por Li e Tang (2008), eles utilizaram uma ferramenta de pesquisa acadêmica[2] visando encontrar um revisor para um determinado estudo chamado “*The boosting approach to machine learning: An overview*”. Ao inserir o título do artigo para consulta na ferramenta, o primeiro resultado obtido foi de um especialista chamado “J. Ross Quinlan”.

Ao analisar a carreira de “J. Ross Quinlan”, porém, os autores constataram que a competência que o qualificava para revisar o artigo em questão não era mais trabalhada por ele em suas publicações desde antes do ano 2000. Os autores do

[2] AMiner: https://aminer.org/.

estudo concluíram que, como o interesse profissional de “Quinlan” havia mudado, ele já poderia não ser apropriado para revisar o artigo buscado. Ou seja, sua competência naquela área já não era evidente.

Para solucionar o problema descrito, segundo os autores, é necessário considerar o fator temporal, pois este afeta diretamente a qualidade do resultado para a consulta realizada. É partindo do mesmo princípio que o presente trabalho foi idealizado. A evidência de que o fator temporal é determinante ao produzir-se conhecimento acerca das áreas de aptidão de um profissional motiva a necessidade de investigar mais profundamente o impacto dessa variável na qualidade de um agrupamento por competências.

3. DESCRIÇÃO DO PROCESSO

Este capítulo descreve o processo de produção da aplicação tCALC, apresentando todas as fases de trabalho executadas desde o início do projeto partindo do CALC até a concepção da versão final da aplicação. Ele evidencia, de maneira paralela, as decisões tomadas e a base teórica utilizada para apoiá-las.

Inicialmente. Apresenta-se o procedimento de mineração de dados adotado, ou seja, as ações que foram tomadas em relação aos dados para gerar um agrupamento natural e que possibilite a extração de informação relevante com consideração de aspecto temporal.

Após, é descrita a implementação propriamente dita, ou seja, as modificações que o código do CALC sofreu e que o fizeram se tornar o tCALC.

Por fim, são trazidos os experimentos realizados sobre o resultado final e a análise da qualidade de seus agrupamentos.

3.1. PROCEDIMENTO

As atividades deste trabalho seguem as etapas definidas pelo processo de KDD (*Knowledge Discovery in Databases* - Descoberta de Conhecimento em Bancos de Dados). Esse processo foi desenvolvido para servir de ferramenta a auxiliar na extração de informações úteis do grande volume de dados digitais produzidos atualmente e, principalmente, de dados que, se considerados individualmente, possuam baixo teor informativo [11].

O KDD é composto por 5 fases que se fazem presentes no procedimento deste trabalho (seleção, pré-processamento, transformação, mineração de dados, interpretação/avaliação) que, embora propostas nessa sequência, são organizadas de maneira diferente conforme as necessidades do projeto, o que não afeta o resultado.

A Figura 1 esquematiza a sequência dos passos do KDD demonstrando as transformações que os dados sofrem até que, em um cenário ideal, o conhecimento possa ser adquirido. As próximas seções vão fazer uma breve revisão bibliográfica de cada etapa e um comentário sobre aplicação dessas ao trabalho.

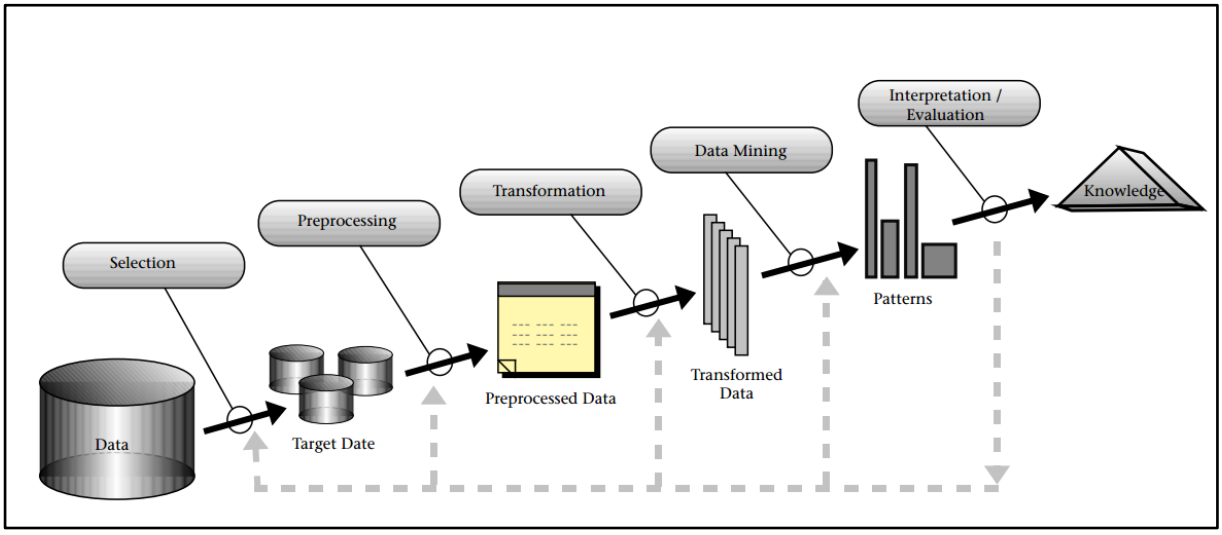

Figura 1 - As Etapas do Processo de KDD.

Fonte: Fayyad et. al., 1996.

3.1.1. Seleção

A primeira etapa que o KDD sugere é a seleção, dentre todo o conjunto de dados disponível, daqueles que são relevantes ou que ao menos haja a suspeita sobre sua relevância para a produção do conhecimento ou a obtenção da informação desejada. Essa etapa também pode levar o analista a produzir novos dados caso isso se mostre necessário ou benéfico para a processo. A inclusão de metadados e todos os tipos de dados – inclusive de fontes externas – que possam oferecer algum suporte semântico aos dados principais também pode beneficiar a descoberta [8].

O primeiro aspecto dessa etapa, em relação ao procedimento deste projeto, foi a definição da informação que se pretende extrair do conjunto de dados sobre os quais será trabalhado. Podemos sintetizá-lo através da primeira pergunta que devemos responder na tomada de decisões: “o que queremos saber sobre os dados?”.

A resposta a essa pergunta herda seus fundamentos do projeto executado anteriormente por Esteves (2015), onde a resposta era “a área de atuação de um dado profissional” ou “quais profissionais atuam na mesma área”. Mas no caso atual, da expansão do tema que está sendo proposta nesta obra, a resposta também se expande: queremos saber dos dados qual a área de atuação de um profissional em um dado momento no tempo; ou ainda: quais profissionais atuaram (ou têm competência) na mesma área em um mesmo momento do tempo.

Nessas respostas tem-se objetivos bem definidos para o KDD. A partir disso, é possível determinar quais dados e atributos podem contribuir para alcançá-los. Esse segundo aspecto também exigiu tomada de decisão e análise da natureza dos dados trabalhados: os currículos Lattes no formato XML. É neste momento que o tCALC começa a se diferenciar do CALC.

Foi decidido inicialmente que tCALC, assim como seu predecessor, deveria selecionar alguns atributos específicos dos documentos XML que contivessem informações sobre competências dos profissionais. Porém, diferentemente do CALC, havia uma necessidade adicional: os atributos selecionados deveriam expressar o fator temporal além da competência, ou seja, deveriam relacionar a atuação do profissional com o momento no tempo onde ela ocorreu. Os atributos selecionados

pelo CALC não continham essa informação necessária. Portanto, foi detectada a necessidade de considerar outros atributos que pudessem contribuir nesse aspecto.

Foi aí que o projeto esbarrou na primeira limitação dos tipos de dados utilizados: até onde constatou-se, os currículos Lattes não possuíam muitos campos que reproduzissem simultaneamente as informações de área e momento de atuação de maneira consistente. Essa situação condicionou a tomada de decisão a ser feita em favor da utilização dos campos referentes à produção bibliográfica dos profissionais, mais precisamente em relação aos artigos publicados. Os atributos presentes nos dados que poderiam satisfazer às necessidades, visto que os artigos que um autor produz refletem a área de pesquisa à qual ele tem se dedicado, são 'TITULO-DO-ARTIGO' e 'ANO-DO-ARTIGO'. Eles foram utilizados para aquisição de dados sobre competência e temporalidade respectivamente.

Como já se esperava, essa escolha mostrou fragilidades desde os primeiros experimentos que seguiram sua implementação. A pequena quantidade de dados disponível nesses dois atributos impactou diretamente na qualidade do agrupamento. Isso porque os lexemas limitados extraídos desses campos não traziam uma visão expandida da área de atuação, mas apenas ideias muito específicas relacionadas a cada uma das publicações selecionadas.

Para aprimorar a mineração em relação ao problema trazido anteriormente, foi proposta a inclusão de dados provenientes de outras bases que pudessem aumentar a expressividade dos dados extraídos de 'TITULO-DO-ARTIGO'. A base de dados escolhida para esse fim foi o *Lattes Expertise Retrieval* (LExR).

O LExR trata-se de um projeto que, para fins de análise de busca acadêmica e recuperação de especialidades, sumarizou em um banco de dados não-relacional informações como título, *abstract*, autores, palavras-chave e área de aproximadamente todos os artigos registrados na Plataforma Lattes [16].

Através dos títulos dos artigos extraídos anteriormente e dessa base de dados, foi possível, para cada artigo, consultar e incluir os dados selecionados para mineração as palavras-chave e áreas extraídas do LExR, ambas informações relevantes para o agrupamento proposto.

3.1.2. Pré-Processamento

É indicado que haja um tratamento adequado dos dados, sejam eles estruturados ou não, antes do processo de mineração, para que os algoritmos que a implementam possam estar imunes a inconsistências. Essas ações podem produzir minerações com grandes vantagens em termos de qualidade [9].

Esse tratamento inclui, entre outras, três técnicas mais comuns: (1) preparação de dados, (2) tratamento de valores faltantes e (3) tratamento de ruídos [9].

No caso do projeto tCALC, as atividades de pré-processamento se limitam à ação conhecida como remoção de *stopwords*. Essa ação é bastante comum em mineração de dados textuais e consiste na remoção daquelas estruturas das linguagens naturais que conectam as ideias dentro de uma frase mas que, por si só, não apresentam muito valor semântico para a mineração. Exemplos comuns desses conectivos são as preposições, artigos e pronomes [10].

Além disso, não foram tomadas ações de pré-processamento visto que os dados trabalhados em relação aos atributos escolhidos: (1) apresentam-se homogêneos, (2) não é comum que hajam valores faltantes e (3) não apresentam muito ruído além das *stopwords*.

3.1.3. Transformação

Após a fase de pré-processamento garante-se que os dados tenham adquirido uma condição consistente, mas não garante que eles estejam prontos para a mineração. A última ação a ser tomada antes de aplicar os algoritmos de mineração deve ser a transformação dos dados de seu formato original em um que os algoritmos de mineração aceitem como entrada, o que costuma variar entre diferentes implementações [9].

Essa transformação também costuma incluir estratégias para redução do volume de dados em favor do desempenho dos algoritmos e possivelmente devido às limitações dos sistemas. Técnicas comuns para esse fim são: (1) redução de dados, (2) discretização e (3) normalização [9].

O tCALC executa a transformação através da redução de dados via eliminação de todo o dado que não diga respeito ao conteúdo dos campos ‘TITULO-DO-ARTIGO’ e ‘ANO-DO-ARTIGO’. Nesse momento, a aplicação cria um arquivo de texto referente a cada currículo contendo apenas esse conteúdo. Após isso, é feita uma consulta no banco de dados referente ao LExR para cada artigo, incluindo no arquivo de texto as palavras-chave e áreas obtidas da base de dados.

Ações como discretização e normalização são dispensadas porque os dados numéricos presentes no atributo “ano do artigo” são descartados a partir a segunda fase de transformação, onde a aplicação usa os anos para segregar o arquivo texto criado anteriormente, dividindo-o em vários arquivos menores, cada qual relativo a um ano e contendo apenas os termos referentes aos títulos dos artigos produzidos naquele ano pelo determinado autor. Esses arquivos, que possivelmente serão vários para um mesmo autor em diferentes anos, são colocados na entrada da fase de mineração. A Figura 2 ilustra o processamento descrito, onde o arquivo criado na primeira fase de transformação é quebrado na segunda fase.

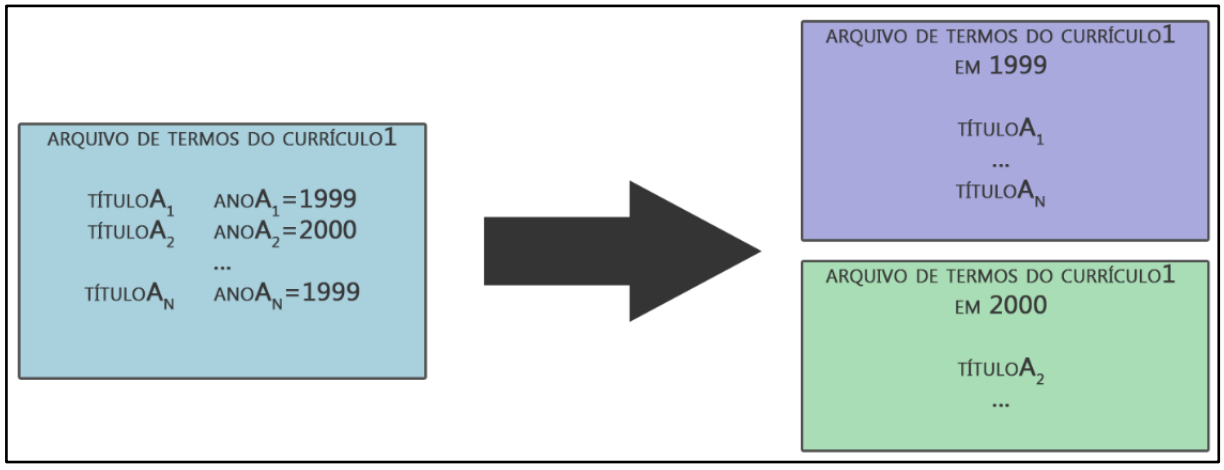

Figura 2 - Abstração da Segregação de Termos pelos Respectivos Anos.

Fonte: autor.

Após isso, foi adicionada a etapa adicional de consulta ao LExR. Nesta etapa, o título de cada um dos artigos foi consultado no banco de dados, os campos *Keywords* (do inglês, palavras-chave) e Áreas foram recuperados e adicionados aos arquivos de termos respectivos. Quando as consultas não conseguem obter esses campos para alguns artigos, a aplicação trabalha apenas com os títulos deles.

A Figura 3 ilustra o processo de transformação combinado com as consultas ao LExR.

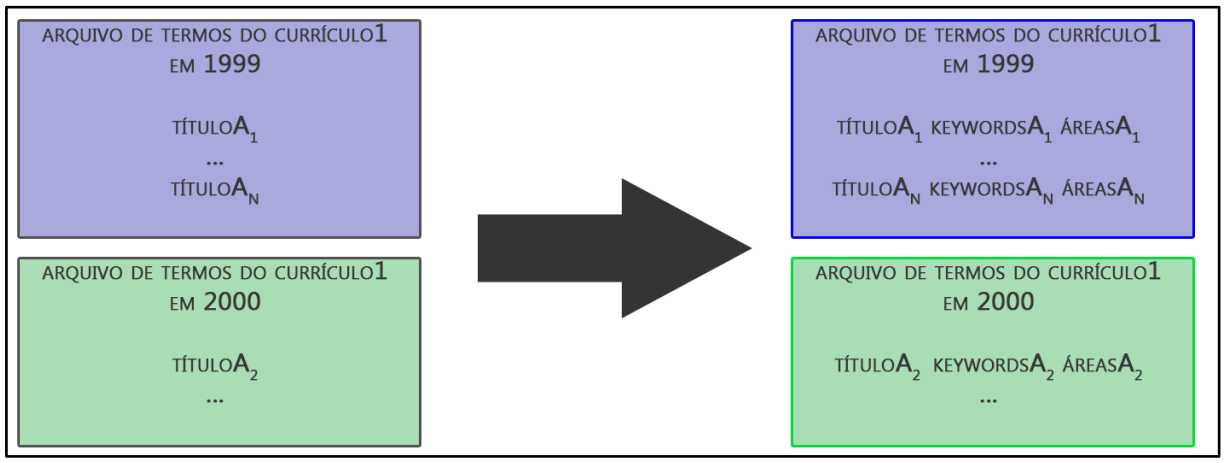

Figura 3 - Abstração da Complementação dos Termos a partir de Consultas ao LExR.

Fonte: autor.

3.1.4. Mineração

Esta é, provavelmente, a etapa mais sofisticada do ponto de vista computacional no KDD. Segundo Fayyad et al (1996), a “mineração de dados é a aplicação de algoritmos específicos para extrair padrões a partir de dados”.

É desejável que esses padrões possuam algumas qualidades tais como validez, utilidade, novidade, entre outras características que alguns autores agruparam e nomearam essa generalização como *interestingness*, ou seja, a qualidade daquilo que é interessante. Os padrões gerados pela etapa de *data mining* devem ser interessantes para serem capazes de oferecer novos conhecimentos ao analista [8].

A mineração pode ser feita de diversas formas, através de diversas técnicas e ainda, cada técnica pode ser implementada por algoritmos diversos e por vezes bem distintos [8]. A técnica abordada no procedimento deste trabalho foi idealizada por Esteves (2015) e consiste no agrupamento de dados usando os algoritmos *BestStar* e *K-medoids* implementados pelo URSA, um *framework* que oferece algoritmos para cálculo de similaridade e agrupamento de dados e que permite a aplicação dessas funcionalidades sobre diversos tipos de dados [1].

O tCALC, a exemplo do CALC, utiliza o algoritmo *BestStar* inicialmente para estimar uma quantidade ótima de *clusters* a serem gerados para o conjunto de dados de entrada. Essa etapa é uma preparação para a execução do *K-medoids* já que esse segundo algoritmo espera que seja fornecida como entrada a quantidade de *clusters* que serão gerados para os dados, enquanto o *BestStar* não precisa dessa informação.

Dessa forma, o *BestStar* é executado anteriormente recebendo como entradas apenas os dados para agrupamento e um limite para os valores de similaridade entre eles. Ele executa a clusterização agrupando currículos cuja similaridade é inferior ao limite estipulado e devolve na sua saída a quantidade de *clusters* gerados para aqueles dados.

Após isso, o tCALC executa o *K-medoids* utilizando a estimativa gerada pelo *BestStar*. Isso é feito porque o *K-medoids* costuma apresentar maior qualidade nos agrupamentos que constrói [1].

3.1.5. Interpretação/Avaliação

A última fase do processo é mais dependente de intervenção humana que as demais. As saídas dos algoritmos da fase anterior são tratadas de forma a evidenciar as informações obtidas através de gráficos ou qualquer outro tipo de representação que seja mais simpática à análise humana [8].

Após a disponibilização visual desses resultados, é necessário que o analista os avalie e compare na tentativa de obter informações evidenciadas pelos padrões formados ou, mediante a devida constatação, decida pela alteração do processo para uma nova tentativa de KDD [8].

Neste projeto, a aplicação gera como saída do agrupamento uma hierarquia de diretórios no sistema de arquivos que revelam os *clusters* formados. Essa árvore de diretórios se organiza com uma raiz chamada *Clusters*, dentro da qual existem pastas referentes a cada período de tempo considerado no agrupamento e denominadas com base no primeiro ano do intervalo considerado (por exemplo, a pasta referente ao triênio 2010-2012 é denominada 2010). Dentro da pasta de cada ano estão diretórios numerados referentes aos *clusters* gerados nos quais estão os currículos Lattes atribuídos a esses grupos.

Para ficar mais clara a saída do tCALC, um exemplo é o seguinte: após o agrupamento de um conjunto de currículos localizado no diretório “Currículos” do sistema operacional, o tCALC vai gerar uma pasta com caminho “Currículos\Clusters”. Para este exemplo, se todas as publicações dos currículos foram elaboradas nos anos 2000, 2001 e 2003, a aplicação cria mais 3 diretórios com caminhos “Currículos\Clusters\2000”, “Currículos\Clusters\2001” e “Currículos\Clusters\2003”. Para cada um desses 3 períodos é feito um agrupamento que gera pastas de *cluster* dentro dos respectivos diretórios de ano.

As análises deste projeto foram feitas com base na saída descrita acima a fim de gerar deduções sobre a qualidade dos *clusters* gerados e da possibilidade de aprimorar o processo para melhorar os resultados.

3.2. IMPLEMENTAÇÃO

Esta seção apresenta as modificações em relação ao código original do CALC que contribuíram para produzir o estado atual do tCALC, a3ssim como as ideias e propostas que motivaram essas modificações.

3.2.1. Classe *ManipuladorXML*

A função principal dos métodos dessa classe é percorrer todos os arquivos XML de currículos, recuperar os valores dos atributos escolhidos para o cálculo de similaridade da mineração e criar estruturas de dados – denominadas por Esteves (2015), autor do CALC, como documentos – contendo esses valores denominados termos.

A função da classe permaneceu inalterada no desenvolvimento do tCALC, porém a maneira como ela é implementada mudou drasticamente devido ao uso do XPath – linguagem de consulta compatível com Java que auxilia na navegação pelos nodos de um documento XML – para facilitar o atendimento às necessidade do tCALC e a integração do projeto com a base de dados LExR.

As novidades em relação ao código original dizem respeito a: (1) a mudança dos atributos pesquisados, (2) a criação dos documentos e (3) uso da interface ConsultorBD. Ainda vale mencionar o uso do XPath, API (*Application Programming Interface)* que auxilia na manipulação de documentos XML, que promoveu uma mudança severa no código da classe ManipuladorXML.

Sobre (1), enquanto o CALC buscava os campos contendo os prefixos ou sufixos ‘NOME-DA-ESPECIALIDADE’, ‘AREA-DO-CONHECIMENTO’, ‘TITULO’, ‘TEXTO-RESUMO-CV-RH’ e ‘PALAVRA-CHAVE’, o tCALC se interessa por apenas dois campos: ‘TITULO-DO-ARTIGO’ e ‘ANO-DO-ARTIGO’.

Acerca de (2), o CALC cria um documento único onde ele coloca todos os valores encontrados para os atributos listados acima. O tCALC, por outro lado, cria vários documentos onde cada um corresponde a um dos diferentes valores encontrados para ‘ANO-DO-ARTIGO’ e o conteúdo desses documentos é os

conteúdos de ‘TITULO-DO-ARTIGO’ correlatos com o intervalo de tempo definido junto à interface gráfica que inicia naquele ano.

E sobre a terceira e última mudança, o ManipuladorXML utiliza-se de um objeto ConsultorBD para abrir conexão com o banco de dados LExR, consultar os valores adicionais de palavras-chave e áreas contidos na base para cada artigo visitado e fechar a conexão.

3.2.2. Classe *Calc*

Essa classe tem papel central na aplicação. É ela que faz a chamada de todos os métodos que implementam as quatro primeiras etapas do KDD.

A primeira modificação em relação ao código original tem impacto direto na etapa de transformação dos dados. O CALC, para gerar os arquivos contendo todos os termos que ele usa no cálculo de similaridade e oriundos dos diversos atributos que ele considera, chama o ManipuladorXML referido anteriormente para criar as listas de documentos com termos e, depois, simplesmente cria um arquivo de texto para cada documento desses. O tCALC, por sua vez, precisa tratar as mudanças que o ManipuladorXML sofreu e separar os documentos de termos que agora estão particionados por anos e criar arquivos para eles em diretórios referentes a cada ano. A Figura 3 ilustra a diferença entre os processos.

A segunda modificação foi feita para acompanhar a mudança da abordagem de mineração. Enquanto o CALC só executava os algoritmos de agrupamento uma vez para todos os dados, o tCALC precisa executá-los *M* vezes, para *M* sendo a quantidade de anos nos quais pelo menos um artigo foi escrito pelos autores dos currículos amostrados. Então foi criado um laço no Calc para performar essas chamadas iteradamente, criando um diretório de saída correspondente a cada período para o qual foi feito agrupamento e colocando os *clusters* nesses diretórios.

A última modificação no código deu-se em decorrência de um erro na chamada do algoritmo *K-medoids* que não estava de acordo com as novas exigências do código. Anteriormente o CALC não colocava um limite de iterações para o *K-medoids*. O tCALC, devido à natureza diferente de seus dados de entrada, estava encontrando problemas onde o algoritmo parecia divergir, nunca atingindo

um valor ótimo para os centróides. Foi adotada, em decorrência disso, uma nova política onde o valor do centróide é fixado após um máximo de 50 iterações, podendo apresentar um valor sub-ótimo.

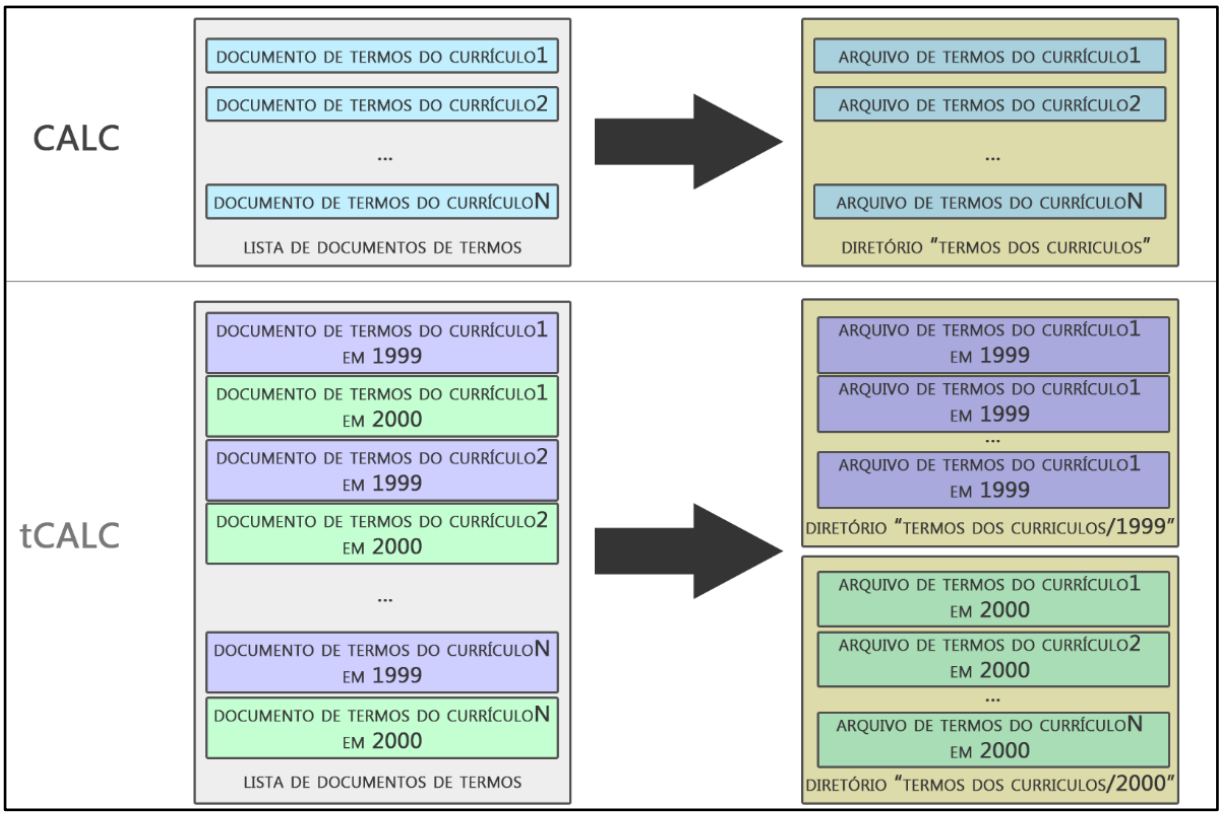

Figura 4 - Comparativo da Criação de Arquivos de Termos em CALC e tCALC.

Fonte: autor.

3.2.3. Classe *TextFilesUtilityCALC*

A mudança ao código desta classe foi bastante sutil. A classe em questão é responsável por criar o objeto de dados que será usado pelo agrupador para o cálculo de similaridade entre os currículos, ou seja, é uma das classes responsáveis pela etapa de transformação dos dados. A alteração reflete as mudanças feitas em outras partes do programa e descritas anteriormente sobre os nomes dos arquivos de texto que concentram, depois do pré-processamento dos dados, os termos de interesse para comparação.

Em *TextFilesUtilityCALC*, de maneira inversa ao que ocorre na classe *Calc*, faz-se necessário obter os nomes dos arquivos XML dos currículos a partir dos nomes dos arquivos TXT de termos para que a clusterização consiga resgatar esses arquivos XML para entregar na saída da aplicação, conforme explicado anteriormente.

Na implementação de CALC, a diferença entre esses nomes de arquivos se dava apenas na extensão deles: enquanto os originais eram .XML, os arquivos de termos eram .TXT. Já no tCALC, com a inserção do ano no nome dos arquivos de termos, foi necessário uma linha adicional no código da classe para retirar os caracteres referentes a essa inserção. Para evidenciar a necessidade da abordagem diferente descrita acima, a Figura 1 ilustra como era feita antes a conversão de nomes e como é feita no tCALC. Além da mudança reversa da extensão, os 7 primeiros caracteres referentes aos parênteses, ano e espaço em branco devem ser retirados para voltar ao nome original.

Figura 5 - Exemplo comparativo das transformação de nomes de currículos XML em nomes de arquivos TXT de termos.
Fonte: autor.

3.2.4. Classe *ClusteringUtilityCALC*

Essa classe é responsável pela criação de arquivos e diretórios referentes ao processo de *clustering*. Ela cria, por exemplo, o diretório de cada *cluster* onde os currículos XML correspondentes são colocados após a mineração.

No CALC, os diversos *clusters* se localizavam em uma mesma pasta chamada Clusters, já que os grupos naturais eram referentes ao conjunto de dados completo. As mudanças do tCALC, porém, exigiram que fosse criada uma pasta para cada ano, onde os *clusters* referentes àquele ano pudessem se localizar. É na classe ClusteringUtilityCALC que essa mudança foi implementada.

3.2.5. Classe JanelaPrincipal

A classe JanelaPrincipal, como o nome sugere, define a aparência e as interações da interface gráfica da aplicação. A mudança feita nela é bastante discreta, embora necessária para o tCALC. Foi adicionado o campo “Quantidade de anos do Clustering” que determina através de sua entrada a quantidade de anos do período considerado para clusterização, ou seja, se a entrada for 3, por exemplo, como na Figura 5 que mostra o comparativo entre as interfaces do CALC e tCALC, a aplicação agrupa currículos que tiveram publicações semelhantes no mesmo triênio.

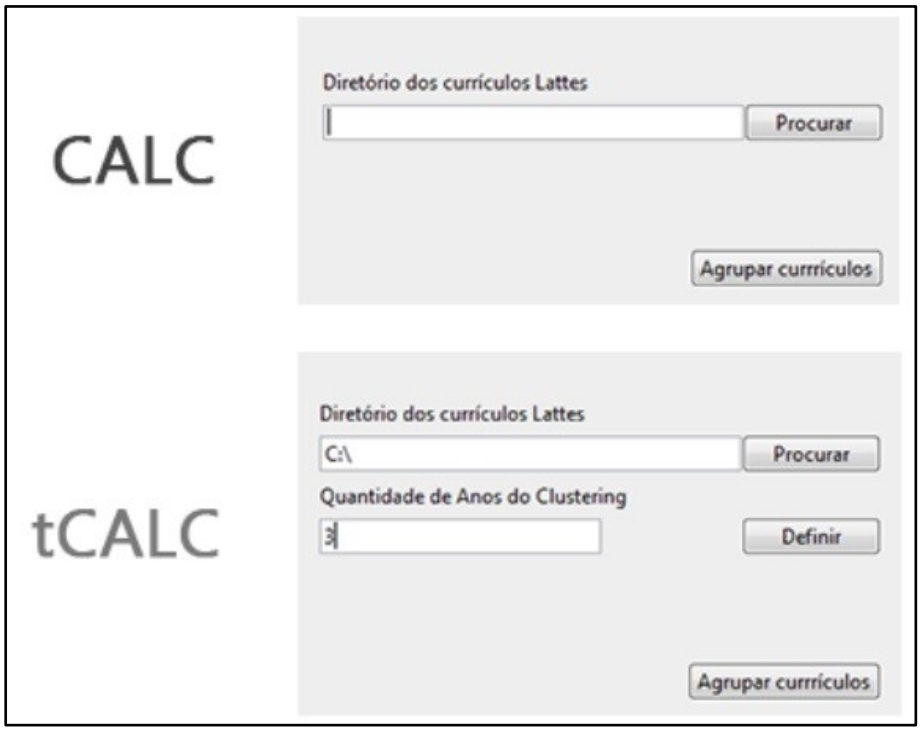

Figura 6 - Comparativo entre as interfaces gráficas de CALC e tCALC.
Fonte: autor.

3.2.6. Interface ConsultorBD

A interface ConsultorBD e as classes que a implementam, ConsultorMongo e ConsultorMySQL, foram elaboradas para dar suporte à comunicação da aplicação com a base de dados LExR.

Inicialmente foi implementada apenas a classe ConsultorMongo, visto que seria usado um servidor MongoDB para comportar os dados do LExR. Essa classe possuía três métodos: um para abrir conexão com o banco, outro para a consulta propriamente dita e um último para fechar conexão.

Após alguns experimentos utilizando MongoDB apresentarem uma latência muito alta nas consultas, decidiu-se experimentar o MySQL em busca de melhor desempenho. Adicionalmente, essa decisão foi tomada por conta da vantagem do MySQL em lidar com consultas insensíveis a acentos, visto que havia uma inconsistência entre os acentos dos títulos do artigos nos currículos Lattes e no LExR, o que inviabilizava as consultas se elas não fossem capazes de corrigir automaticamente essa diferença.

Para essa mudança, os três métodos implementados pelo ConsultorMongo foram abstraídos para a interface ConsultorBD – a qual ConsultorMongo passou a implementar – e implementados pela classe ConsultorMySQL com os mesmo propósitos adaptados para o novo banco.

Os experimentos com MySQL apresentaram melhorias de desempenho conforme esperado, logo a versão final do tCALC utiliza essa abordagem para consultar o LExR. Apesar disso, a classe ConsultorMongo permanece no projeto visando oferecer essa opção a desenvolvedores que forem trabalhar sobre esse código possivelmente.

3.3. EXPERIMENTOS

3.3.1. Experimentos Preliminares

A amostra de currículos utilizada possui 206 arquivos contendo currículos Lattes no formato XML de pesquisadores de programas de pós-graduação da UFSC das áreas: agronomia, aquicultura, bioquímica, ciência da computação, ciências humanas, enfermagem, engenharia de alimentos, engenharia civil, engenharia mecânica, farmácia, farmacologia, filosofia, física, história, literatura e química. Essa amostra foi a mesma utilizada na implementação do CALC e foi escolhida pelo autor daquele projeto de forma a possuir elementos de áreas distintas mas também de áreas com alguma semelhança para analisar efeitos que mudanças nos parâmetros causariam aos resultados do agrupamento [1].

Os currículos da amostra são identificados pelos nomes das áreas aos quais pertencem, seguidos de números para diferenciar uns dos outros (por exemplo, Agronomia 3 ou Aquicultura 1). Alguns outros currículos, pertencentes a profissionais de Ciência da Computação, são denominados com os nomes dos profissionais que retratam. Essas estratégias foram tomadas para permitir que, na etapa de Análise, seja possível distinguir quais *clusters* possuem currículos semelhantes de fato [1].

Os experimentos preliminares visaram garantir a efetividade e a eficiência preliminar do código. Eles foram elaborados com um conjunto limitado de dados.

O sub-conjunto de teste possui 11 dos currículos acima e foi tomado para verificar o funcionamento completo do agrupamento de forma a agilizar os testes de código. À medida que a implementação apresentou efetividade no agrupamento desses 11 currículos, ela foi aplicada ao grupo maior de 206 para que se avaliasse a qualidade do agrupamento, tomando como comparativo os resultados alcançados no projeto CALC.

Como mencionado anteriormente, não foi objetivo do projeto no seu primeiro momento que se obtivesse resultados ideais ou execução em tempo ótimo. A preocupação inicial era que o sistema respondesse corretamente ao agrupamento utilizando-se da métrica de competências e considerando fortemente o fator temporal.

A execução dupla dos algoritmos *BestStar* e *K-medoids* não estava sendo possível devido a erros apresentados na execução do segundo algoritmo e oriundos de inconsistências entre algumas partes do código do tCALC e as exigências para plena execução do *K-medoids.* Nesse momento foi decidido que essas inconsistências seriam tratadas futuramente no projeto e que seria feita apenas a execução simples do *BestStar* a fim de testar o funcionamento das outras partes da implementação.

Os testes realizados buscaram fazer o agrupamento apenas de currículos que possuíam obras publicadas no mesmo ano. Dessa forma, para a amostra de 206 currículos, foram gerados 48 agrupamentos diferentes, cada um respectivo a um ano dentro do intervalo de 1962 e 2016.

Os resultados da execução sob esses termos não apresentaram qualidade de agrupamento, como esperado: em um cálculo da média aritmética de *clusters* que apresentaram mais de um elemento em relação à quantidade total de *clusters* gerados pelo algoritmo, ficou constatado que apenas 3% dos grupos se encaixavam nessa descrição.

O algoritmo levou 197 segundos para performar todos os agrupamentos, muito mais tempo que os 14 segundos do CALC. Admite-se que seja natural que o tCALC nesses experimentos tenha sido mais lento que CALC porque ele se propõe a executar muito mais iterações sobre os dados, visto que executa o algoritmo de agrupamento 48 vezes para o experimento realizado: uma para cada ano amostrado. Já o CALC executa apenas uma vez o algoritmo sobre um conjunto maior de dados.

3.3.2. Experimentos com *K-medoids*

Nesta fase os experimentos foram elaborados buscando avaliar a qualidade do agrupamento e, para isso, utilizou-se o lote de amostras com 206 currículos.

Após mais avaliações do código, foram detectadas as falhas que estavam causando o não funcionamento do *K-medoids*: os centróides não estavam convergindo para um número ilimitado de iterações do algoritmo e a quantidade de *clusters* não estava sendo corretamente estimada após a execução do *BestStar.*

Depois de corrigidas essas falhas, o agrupamento teve sua qualidade melhorada grandemente. O problema dos *clusters* individuais foi superado e já se observou algum sentido nos agrupamentos: profissionais que se sabia previamente pertencer a uma mesma área já estavam sendo colocados sob um mesmo *cluster*.

Pode-se visualizar isso no próximo experimento realizado, considerando triênios ao invés de anos isolados. Para o exemplo abaixo, foram tomados alguns *clusters* dos 11 gerados para o triênio que vai de 2011 a 2013. Observando-se que os currículos da amostra foram denominados com suas grandes áreas e numerados para fim de controle com exceção daqueles de Ciência da Computação que recebem o nome de seus proprietários [1].

- O primeiro *cluster* incluiu os currículos: Bioquímica 6, Agronomia 3, Farmácia 11, Farmácia 12, Farmacologia 6, História 9.
  Podemos perceber que profissionais de áreas relativamente próximas (Bioquímica, Farmácia e Farmacologia) foram agrupados juntos. O currículo de História é possivelmente um *outlier*, enquanto é provável que o de Agronomia possua certa proximidade das demais áreas.
- Segundo *cluster*: Eng. Mecânica 2, Eng Mecânica 12, Filosofia 4, Filosofia 6, História 11, História 13, História 6, História 8, Literatura 10, Literatura 4, Literatura 5.
  Neste agrupamento, apesar dos *outliers* (Eng. Mecânica) também podemos observar currículos de áreas semelhantes relacionadas às Ciências Humanas sendo trazidos juntos.
- Terceiro *cluster*: Filosofia 3, Ciências Humanas 6, Filosofia 2, Literatura 7.
  Esse cluster se demonstra organizado, com todos os currículos pertencendo a áreas próximas das Ciências Humanas.

3.3.3. Experimentos com LExR Integrado

Os experimentos desta etapa foram bastante semelhantes aos anteriores, visto que a única mudança – embora substancial – foi a adição dos dados da base de dados externa LExR no processo. Com os dados adicionais de palavras-chave e

áreas dos artigos do Lattes, foi possível aumentar a expressividade dos dados de entrada.

Se for feita uma análise dos *clusters* gerados para o mesmo triênio trazido no exemplo anterior à integração com o LExR, considerando publicações feitas entre 2011 e 2013, já se percebe imediatamente uma quantidade maior de *clusters*: 14 contra os 11 obtidos anteriormente.

Tomando como ponto de partida um currículo qualquer pertencente a cada um dos três grupos mostrados no exemplo anterior, três dos novos *clusters* gerados são:

- *Cluster* contendo o currículo Farmácia 11: Aquicultura 11, Bioquímica 1, Bioquímica 2, Bioquímica 4, Bioquímica 6, Bioquímica 7, Bioquímica 8, Farmácia 9, Farmácia 11, Farmácia 12, Farmacologia 11, Química 2, Química 8.

  Este grupo apresenta-se organizado em torno das áreas de Farmácia, Bioquímica, Farmacologia e Química, áreas possivelmente próximas. Pode-se observar apenas um *outlier* (Aquicultura 11).
- *Cluster* contendo o currículo Eng. Mecânica 2: Agronomia 2, Eng. Alimentos 4, Eng. Civil 5, Eng. Civil 8, Eng. Civil 9, Eng. Civil 11, Eng. Mecânica 1, Eng. Mecânica 2, Eng. Mecânica 3, Eng. Mecânica 8, Eng. Mecânica 13, Física 6, Raul - 09-04, Ricardo - 10-04, Silvia - 14-04.
  Esse grupo demonstrou comportamento bastante peculiar, apresentando múltiplos subgrupos bem definidos de currículos: um subgrupo de currículos da área de Engenharia Civil, outro da Engenharia Mecânica e outro de Ciência da Computação. Esse fenômeno pode ser explicado pela presença de um centróide simetricamente próximo de todos esses subgrupos ou ainda pela limitação do cálculo da similaridade devido à escassez dos dados.
- O currículo Filosofia 3 se encontra no *cluster* mais distinto do período analisado. Por motivo de clareza do texto, o grupo não é apresentado neste parágrafo, mas ele pode ser conferido no Apêndice B sob a denominação “Cluster 14”. O comportamento dele se assemelha ao descrito no parágrafo acima, onde manifestam-se diversos subgrupos dentro do mesmo *cluster*.

Os resultados acima não foram colocados com o intuito de comprovar a qualidade do agrupamento, mas apenas para mostrar a evidência e descrever

aproximadamente a dinâmica dos experimentos realizados. O próximo capítulo se preocupa em avaliar mais detalhadamente a qualidade.

4. CONCLUSÃO

4.1. ANÁLISE DE RESULTADOS

A análise dos resultados deste projeto dá-se de forma puramente argumentativa com base nos experimentos apresentados nos Apêndices A, B, C e D que acompanham este texto.

Como mencionado anteriormente, os currículos fornecidos como entrada da aplicação foram nomeados com as áreas dos profissionais que eles representam com o intuito de facilitar a observação de *clusters* gerados com sucesso ou não. A argumentação presente nestas análises parte da definição que as seguintes áreas presentes no conjunto de entrada são mutuamente próximas no ponto de vista da engenharia de conhecimento:

- Química, Bioquímica, Farmácia e Farmacologia;
- Agronomia e Aquicultura;
- Ciências Humanas, Filosofia, História e Literatura;
- Física e Engenharia Civil;
- Física e Engenharia Mecânica;
- Física e Química;
- Química, Física e Engenharia de Alimentos.

Outras relações não listadas podem ser verdadeiras mas não foram consideradas nas análises a seguir.

As análises foram feitas para todos os *clusters* gerados nos experimentos e levaram em consideração a taxa de agrupamentos bem-sucedidos – ou seja, aqueles que colocam em um mesmo grupo currículos de áreas próximas conforme a lista acima – e, para quando ocorrer, a quantidade de subgrupos formados no *cluster*. Os resultados com ou sem a integração com LExR são comparados ao final de cada experimento.

4.1.1. Experimento 1

Os registros dos agrupamentos que compõem este experimento são trazidos nos Apêndices A e B. Os dados abaixo dizem respeito ao triênio entre 2011 a 2013.

Os dados obtidos para a implementação sem uso do LExR:

- Quantidade total de currículos: 165;
- Quantidade total de *clusters*: 11;
- Quantidade de currículos agrupados corretamente: 127;
- Quantidade de currículos agrupados incorretamente: 38;
- Quantidade de subgrupos: 10 divididos entre 4 grupos;
- Taxa de sucessos: 77%;
- Taxa de fracassos: 23%.

Os dados obtidos para a execução integrada ao LExR são:

- Quantidade total de currículos: 165;
- Quantidade total de *clusters*: 14;
- Quantidade de currículos agrupados corretamente: 142;
- Quantidade de currículos agrupados incorretamente: 23;
- Quantidade de subgrupos: 16 divididos entre 7 grupos;
- Taxa de sucessos: 86%;
- Taxa de fracassos: 14%.

Portanto, com base nas variáveis enumeradas acima, pode-se concluir que há uma melhora com o uso do LExR no experimento. A taxa de sucessos de 86% é considerada satisfatória visto as limitações apresentadas pela base de dados tais como a escassez de termos e campos com informação temporal. A grande quantidade de subgrupos evidencia que, se fosse assumido um número maior de grupos, esses subgrupos se manifestariam como grupos por si só, o que indica que a quantidade ótima de grupos deve ser maior que a escolhida pelo algoritmo. Isso pode abrir espaço para otimizações no processo.

4.1.2. Experimento 2

A exemplo do Experimento 1, os registros deste experimento são trazidos nos Apêndices C e D. Este teste foi elaborado com período mais amplo de agrupamento para avaliar o comportamento do sistema diante do aumento dessa variável. Os dados abaixo expressam o agrupamento feito para o período de 7 anos entre 2009 e 2015.

Os dados obtidos para a implementação sem uso do LExR:

- Quantidade total de currículos: 158;
- Quantidade total de *clusters*: 20;
- Quantidade de currículos agrupados corretamente: 124;
- Quantidade de currículos agrupados incorretamente: 34;
- Quantidade de subgrupos: 8 divididos entre 4 grupos;
- Taxa de sucessos: 78,5%;
- Taxa de fracassos: 21,5%.

Os dados obtidos para a execução integrada ao LExR são:

- Quantidade total de currículos: 158;
- Quantidade total de *clusters*: 24;
- Quantidade de currículos agrupados corretamente: 135;
- Quantidade de currículos agrupados incorretamente: 23;
- Quantidade de subgrupos: 11 divididos entre 5 grupos;
- Taxa de sucessos: 85,4%;
- Taxa de fracassos: 14,6%.

Este experimento também apresentou melhoras após o uso do LExR e apresentou também um número elevado de subgrupos, reforçando as conclusões tiradas no Experimento 1 em relação a essa métrica.

4.2. CONSIDERAÇÕES FINAIS

A aplicação produzida juntamente com esta obra se preocupou, a exemplo de sua antecessora, em oferecer a solução a uma demanda inerente à proposta da plataforma Lattes, o que literatura chama de *expertise retrieval* ou recuperação de especialidades em uma tradução livre. Além disso, o projeto em si buscou apresentar a importância da mineração de dados, da descoberta de conhecimento e da análise de competências.

Até esses requisitos apresentados anteriormente, o projeto tCALC apenas expandiu o que já havia sido abordado no CALC. A diferença fundamental e que permitiu que um novo escopo fosse trabalhado – o aspecto temporal – foi abordada com protagonismo compatível com a relevância dela para o problema apresentado. Ao optar-se por levar em consideração o tempo no *expertise retrieval*, a proposta muda radicalmente tal como os resultados atingidos. E isso foi o que este projeto se preocupou em deixar claro através de seus capítulos de experimentos e análises.

Ainda há muito o que ser produzido pela comunidade científica nas áreas às quais este trabalho pertence. Ainda assim, a expectativa é de que, de alguma forma, esta obra tenha contribuído, seja para gerar conhecimento acerca do assunto ou motivação para que outros projetos avancem cada vez mais o estado da arte.

4.3. TRABALHOS FUTUROS

Durante a execução deste projeto, surgiram algumas ideias relacionadas à aplicação que não puderam ser incluídas em seu escopo. Esta seção se preocupa em enumerá-las. São elas:

- **Aprimoramento do processo de KDD**. Principalmente da etapa de pré-processamento, onde *stopwords* podem ser tratadas de maneira mais efetiva e ainda outras técnicas possam ser empregadas.
- **Otimização temporal da execução**. A execução das diversas etapas do KDD levam um tempo relativamente grande. Separando as etapas para reuso, aprimorando a integração com o banco de dados ou ainda empregando diferentes estruturas de dados e técnicas de iteração podem trazer maior eficiência.
- **Refino dos dados de entrada**. Existem técnicas avançadas de mineração de dados textuais e processamento de linguagem natural – como uso de dicionário de sinônimos, aprimoramento da remoção de *stopwords*, correção ortográfica e tratamento de múltiplos idiomas – que poderiam ser aplicadas para qualificar ainda mais o agrupamento.
- **Reunir CALC e tCALC como dois modos de operação de uma mesma aplicação**, adicionando-se ainda mais utilidades e criando-se um agrupador de currículos Lattes versátil e adaptado a diversas exigências.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] B. F. ESTEVES. *CALC: Agrupamento de Perfil Científico por Afinidade de Áreas de Conhecimento Utilizando Currículos Lattes*. Florianópolis: Universidade Federal de Santa Catarina. 2015.

[2] M.J. ZAKI, and W. MEIRA. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. 1st ed. New York: Cambridge UP, 2014.

[3] E.A.M. MORAIS, and A.P.L. AMBRÓSIO. *Mineração de Textos.* INF - Universidade Federal de Goiás, 2007.

[4] Y. LI, and J. TANG. *Expertise Search in a Time-varying Social Network*. Beijing: Tsinghua University, 2008.

[5] O. ALONSO, and M. GERTZ. *Clustering of Search Results using Temporal Attributes*. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06). ACM, New York, NY, USA, p.597-598.

[6] C. BUCKLEY, and E.M. VOORHEES. *Retrieval Evaluation with Incomplete Information*. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04). ACM, New York, NY, USA, p.25-32.

[7] L. BACKSTROM, D. HUTTENLOCHER, and J. KLEINBERG. *Group Formation in Large Social Networks: Membership, Growth, and Evolution*. Cornell University, 2006.

[8] U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH, and R. UTHURUSAMY. *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA 1-34. 1996.

[9] S. GARCÍA, J. LUENGO, and F. HERRERA. *Data Preprocessing in Data Mining.* In *Intelligent Systems Reference Library vol. 72.* New York: Springer, 2015.

[10] S. VIJAYARANI, J. ILAMATHI, and NITHYA. *Preprocessing Techniques for Text Mining - An Overview*. International Journal of Computer Science & Communication Networks, Vol 5(1), 7-16. 2011.

[11] A. ADHIKARI, and J. ADHIKARI. *Advances in Knowledge Discovery in Databases.* In *Intelligent Systems Reference Library vol. 79.* New York: Springer, 2015.

[12] R. BALANCIERI, A.B. BOVO, V.M. KERN, R.C.S. PACHECO, and R.M. BARCIA. *A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes.* In *Ciência da Informação*. v.34, n. 1, p.64-77. Instituto Brasileiro de Informação em Ciência e Tecnologia. Brasília, 2005.

[13] C.R. SHALIZI. *Advanced Data Analysis from an Elementary Point of View*. New York: Cambridge UP. 2016.

[14] CNPQ. *Sobre a Plataforma*. Disponível em <http://lattes.cnpq.br/>. Acesso em: 15 de jun. 2016.

[15] P.S. YU, X. LI, and B. LIU. *Adding the Temporal Dimension to Search - A Case Study in Publication Search.* The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). Watson Res. Center, Hawthorne, NY, USA. 2005.

[16] V. MAGARAVITE, R.L.T. SANTOS, I.S. RIBEIRO, M.A. GONÇALVES, A.H.F. LAENDER. *The LExR Collection for Expertise Retrieval in Academia.* Department of Computer Science, Universidade Federal de Minas Gerais. Belo Horizonte, MG, Brazil.

APÊNDICE A - Experimento 1 sem Integração ao LExR

Considerando intervalos de 3 anos entre 2011 e 2013.
Período de 2011 - 2013:
Definidos 11 clusters para 165 currículos.

Cluster 1:
- Bioquímica 6.txt
- Agronomia 3.txt
- Farmacia 11.txt
- Farmacia 12.txt
- Farmacologia 6.txt
- História 9.txt

Cluster 2:
- Eng Mecânica 2.txt
- Eng Mecânica 12.txt
- Filosofia 4.txt
- Filosofia 6.txt
- História 11.txt
- História 13.txt
- História 6.txt
- História 8.txt
- Literatura 10.txt
- Literatura 4.txt
- Literatura 5.txt

Cluster 3:
- Filosofia 3.txt
- Ciências Humanas 6.txt
- Filosofia 2.txt
- Literatura 7.txt

Cluster 4:
- Farmacologia 3.txt
- Aquicultura 11.txt
- Bioquímica 3.txt
- Farmacologia 8.txt
- Física 6.txt
- Literatura 9.txt
- Patricia Plentz - 07-04.txt

Cluster 5:
- Bioquímica 2.txt
- Aquicultura 10.txt
- Aquicultura 3.txt
- Bioquímica 1.txt
- Bioquímica 4.txt
- Bioquímica 7.txt
- Bioquímica 8.txt
- Farmacia 9.txt
- Farmacologia 1.txt
- Farmacologia 11.txt
- Farmacologia 2.txt
- Farmacologia 4.txt
- Farmacologia 7.txt
- Literatura 3.txt
- Química 10.txt
- Química 2.txt
- Química 8.txt

Cluster 6:
- Física 5.txt
- Agronomia 4.txt
- Aquicultura 5.txt
- Aquicultura 6.txt
- Eng Alimentos 1.txt
- Eng Alimentos 10.txt

Eng Alimentos 11.txt
Eng Alimentos 12.txt
Eng Alimentos 2.txt
Eng Alimentos 3.txt
Eng Alimentos 5.txt
Eng Alimentos 6.txt
Eng Alimentos 7.txt
Eng Alimentos 9.txt
Eng Mecânica 11.txt
Eng Mecânica 6.txt
Eng Mecânica 7.txt
Farmacia 2.txt
Farmacia 4.txt
Farmacologia 10.txt
Física 10.txt
Física 11.txt
Física 2.txt
Física 7.txt
Física 8.txt
Química 11.txt
Química 12.txt
Química 13.txt
Química 14.txt
Química 4.txt
Química 5.txt
Química 7.txt
Química 9.txt

Cluster 7:

Farmacia 8.txt
Agronomia 1.txt
Agronomia 2.txt
Aquicultura 1.txt
Aquicultura 2.txt
Aquicultura 4.txt
Aquicultura 9.txt
Eng Civil 1.txt
Farmacia 7.txt
Farmacologia 12.txt
História 12.txt
Raul - 09-04.txt

Cluster 8:

Eng Civil 2.txt
Agronomia 6.txt
Aquicultura 8.txt
Eng Alimentos 4.txt
Eng Alimentos 8.txt
Eng Civil 10.txt
Eng Civil 11.txt
Eng Civil 5.txt
Eng Civil 9.txt
Eng Mecânica 1.txt
Eng Mecânica 13.txt
Eng Mecânica 3.txt
Eng Mecânica 4.txt
Eng Mecânica 5.txt
Eng Mecânica 8.txt
Farmacia 1.txt
Farmacia 6.txt
Guto - 06-04.txt
Luciana Rech - 07-04.txt
Química 3.txt
Ricardo -10-04.txt
Rogerio - 14-04.txt

Cluster 9:

Enfermagem 13.txt
Agronomia 5.txt
Agronomia 7.txt
Agronomia 8.txt

Aldo - 06-04.txt
Ciências Humanas 1.txt
Ciências Humanas 2.txt
Ciências Humanas 5.txt
Ciências Humanas 7.txt
Ciências Humanas 8.txt
Enfermagem 1.txt
Enfermagem 10.txt
Enfermagem 11.txt
Enfermagem 12.txt
Enfermagem 14.txt
Enfermagem 2.txt
Enfermagem 3.txt
Enfermagem 4.txt
Enfermagem 5.txt
Enfermagem 6.txt
Enfermagem 7.txt
Enfermagem 8.txt
Enfermagem 9.txt
Eng Civil 6.txt
Eng Civil 8.txt
Filosofia 10.txt
Filosofia 5.txt
Filosofia 8.txt
História 14.txt
História 2.txt
História 5.txt
Literatura 11.txt
Literatura 12.txt
Ricardo S - 14-04.txt
Silvia - 14-04.txt

Cluster 10:

Ronaldo - 14-04.txt
Carina - 06-04.txt
Farmacia 3.txt
Fileto.txt
Física 1.txt
Física 3.txt
Física 4.txt
Física 9.txt
Vania - 14-04.txt

Cluster 11:

west - 9-04.txt
Carla - 06-04.txt
Chris - 09-09.txt
Custodio - 14-04.txt
Eduardo - 14-04.txt
Filosofia 9.txt
Literatura 8.txt
Mario - 09-04.txt
Willrich - 14-04.txt

Total de clusters: 11

Total de clusters com mais de um arquivo: 11

APÊNDICE B - Experimento 1 com Integração ao LExR

Considerando intervalos de 3 anos entre 2011 e 2013.
Período de 2011 - 2013:
Definidos 14 clusters para 165 currículos.

Cluster 1:
- Farmacologia 7.txt
- Bioquímica 3.txt
- Farmacia 3.txt

Cluster 2:
- Enfermagem 7.txt
- Chris - 09-09.txt
- Enfermagem 1.txt
- Enfermagem 3.txt
- Enfermagem 8.txt
- Eng Civil 6.txt
- Filosofia 10.txt

Cluster 3:
- Enfermagem 11.txt
- Agronomia 6.txt
- Agronomia 7.txt
- Agronomia 8.txt
- Aldo - 06-04.txt
- Aquicultura 9.txt
- Ciências Humanas 5.txt
- Ciências Humanas 7.txt
- Enfermagem 10.txt
- Enfermagem 12.txt
- Enfermagem 13.txt
- Enfermagem 14.txt
- Enfermagem 2.txt
- Enfermagem 4.txt
- Enfermagem 5.txt
- Enfermagem 6.txt
- Enfermagem 9.txt
- Filosofia 6.txt

Cluster 4:
- Química 7.txt
- Filosofia 9.txt
- Física 9.txt

Cluster 5:
- Farmacia 8.txt
- Agronomia 1.txt
- Aquicultura 1.txt
- Aquicultura 2.txt
- Aquicultura 4.txt
- Farmacia 6.txt
- Farmacia 7.txt
- Farmacologia 12.txt
- Farmacologia 3.txt
- Farmacologia 4.txt
- História 8.txt

Cluster 6:
- Ciências Humanas 2.txt
- Agronomia 5.txt
- Ciências Humanas 1.txt
- Ciências Humanas 8.txt
- Filosofia 4.txt
- História 11.txt
- História 12.txt
- História 13.txt
- História 14.txt

História 2.txt
História 5.txt
História 9.txt

Cluster 7:

Literatura 11.txt
Filosofia 8.txt
Literatura 10.txt
Literatura 12.txt
Literatura 3.txt
Literatura 8.txt
Literatura 9.txt

Cluster 8:

Física 5.txt
Eng Alimentos 1.txt
Eng Alimentos 12.txt
Eng Alimentos 2.txt
Eng Alimentos 3.txt
Eng Alimentos 7.txt
Eng Mecânica 6.txt
Farmacia 2.txt
Farmacia 4.txt
Farmacologia 10.txt
Física 11.txt
Química 10.txt
Química 11.txt
Química 13.txt

Cluster 9:

Eng Alimentos 10.txt
Agronomia 3.txt
Aquicultura 3.txt
Aquicultura 6.txt
Eng Alimentos 11.txt
Eng Alimentos 5.txt
Eng Alimentos 6.txt
Eng Alimentos 8.txt
Eng Alimentos 9.txt
Eng Civil 1.txt
Farmacologia 1.txt
Farmacologia 2.txt
Farmacologia 6.txt
Farmacologia 8.txt
Química 14.txt

Cluster 10:

Química 9.txt
Aquicultura 10.txt
Aquicultura 5.txt
Aquicultura 8.txt
Eng Mecânica 11.txt
Literatura 7.txt
Química 12.txt
Química 3.txt
Química 4.txt

Cluster 11:

Física 8.txt
Física 10.txt
Física 2.txt
Física 7.txt
Química 5.txt

Cluster 12:

Bioquímica 2.txt
Aquicultura 11.txt
Bioquímica 1.txt
Bioquímica 4.txt
Bioquímica 6.txt
Bioquímica 7.txt
Bioquímica 8.txt
Farmacia 11.txt
Farmacia 12.txt

Farmacia 9.txt
Farmacologia 11.txt
Química 2.txt
Química 8.txt

Cluster 13:

Eng Mecânica 13.txt
Agronomia 2.txt
Eng Alimentos 4.txt
Eng Civil 11.txt
Eng Civil 5.txt
Eng Civil 8.txt
Eng Civil 9.txt
Eng Mecânica 1.txt
Eng Mecânica 2.txt
Eng Mecânica 3.txt
Eng Mecânica 8.txt
Física 6.txt
Raul - 09-04.txt
Ricardo -10-04.txt
Silvia - 14-04.txt

Cluster 14:

Ronaldo - 14-04.txt
Agronomia 4.txt
Carina - 06-04.txt
Carla - 06-04.txt
Ciências Humanas 6.txt
Custodio - 14-04.txt
Eduardo - 14-04.txt
Eng Civil 10.txt
Eng Civil 2.txt
Eng Mecânica 12.txt
Eng Mecânica 4.txt
Eng Mecânica 5.txt
Eng Mecânica 7.txt
Farmacia 1.txt
Fileto.txt
Filosofia 2.txt
Filosofia 3.txt
Filosofia 5.txt
Física 1.txt
Física 3.txt
Física 4.txt
Guto - 06-04.txt
História 6.txt
Literatura 4.txt
Literatura 5.txt
Luciana Rech - 07-04.txt
Mario - 09-04.txt
Patricia Plentz - 07-04.txt
Ricardo S - 14-04.txt
Rogerio - 14-04.txt
Vania - 14-04.txt
west - 9-04.txt
Willrich - 14-04.txt

Total de clusters: 14

Total de clusters com mais de um arquivo: 14

APÊNDICE C - Resumo do Experimento 2 sem Integração ao LExR

Considerando períodos de 7 anos entre 2009 - 2015.
Período de 2009 - 2015:
Definidos 20 clusters para 158 currículos.

Cluster 1:
- Quimica 12.txt
- Farmacia 10.txt

Cluster 2:
- Bioquimica 2.txt
- Bioquimica 1.txt
- Bioquimica 4.txt
- Bioquimica 8.txt
- Farmacia 12.txt
- Farmacologia 1.txt
- Farmacologia 11.txt
- Farmacologia 12.txt
- Farmacologia 3.txt
- Farmacologia 6.txt
- Farmacologia 7.txt
- Farmacologia 8.txt

Cluster 3:
- Farmacologia 2.txt
- Filosofia 11.txt

Cluster 4:
- Enfermagem 9.txt
- Agronomia 6.txt
- Aldo - 06-04.txt
- Ciencias Humanas 7.txt
- Enfermagem 10.txt
- Enfermagem 11.txt
- Enfermagem 12.txt
- Enfermagem 13.txt
- Enfermagem 3.txt
- Enfermagem 4.txt
- Enfermagem 5.txt
- Enfermagem 6.txt
- Filosofia 1.txt
- Filosofia 6.txt

Cluster 5:
- Filosofia 10.txt
- Filosofia 4.txt
- Filosofia 8.txt

Cluster 6:
- Farmacia 7.txt
- Eng Civil 1.txt
- Farmacia 5.txt
- Farmacia 8.txt
- Fisica 1.txt

Cluster 7:
- Agronomia 5.txt
- Ciencias Humanas 5.txt
- Historia 11.txt
- Historia 12.txt
- Historia 4.txt
- Historia 8.txt

Cluster 8:
- Eng Alimentos 12.txt
- Agronomia 4.txt
- Eng Alimentos 1.txt
- Eng Alimentos 4.txt
- Eng Alimentos 7.txt

Eng Alimentos 8.txt
Eng Alimentos 9.txt
Eng Civil 5.txt
Eng Civil 7.txt
Eng Mecanica 6.txt
Lau - 09-04.txt

Cluster 9:

Enfermagem 2.txt
Enfermagem 1.txt
Enfermagem 14.txt
Eng Civil 4.txt
Literatura 11.txt

Cluster 10:

Ciencias Humanas 2.txt
Ciencias Humanas 1.txt
Ciencias Humanas 8.txt
Historia 13.txt
Historia 15.txt
Historia 2.txt
Literatura 5.txt
Literatura 6.txt

Cluster 11:

Literatura 2.txt
Filosofia 9.txt
Historia 7.txt
Literatura 7.txt

Cluster 12:

Literatura 4.txt
Elder - 14-04.txt
Eng Mecanica 12.txt
Fisica 4.txt
Historia 9.txt
Literatura 8.txt

Cluster 13:

Eng Civil 11.txt
Eng Civil 12.txt
Eng Civil 2.txt
Eng Mecanica 1.txt
Eng Mecanica 3.txt
Eng Mecanica 4.txt
Eng Mecanica 5.txt
Farmacia 1.txt
Filosofia 5.txt
Fisica 3.txt
Luiz - 09-04.txt
Mauro - 09-04.txt

Cluster 14:

Quimica 11.txt
Bioquimica 7.txt
Fisica 10.txt
Quimica 1.txt
Quimica 10.txt
Quimica 13.txt
Quimica 2.txt
Quimica 3.txt

Cluster 15:

Fisica 7.txt
Eng Alimentos 2.txt
Eng Alimentos 3.txt
Eng Alimentos 5.txt
Eng Civil 9.txt
Fisica 11.txt
Fisica 5.txt
Fisica 8.txt
Quimica 14.txt
Quimica 5.txt
Quimica 7.txt

Cluster 16:

Farmacologia 10.txt
Eng Alimentos 10.txt
Eng Alimentos 11.txt
Eng Alimentos 6.txt
Eng Mecanica 11.txt
Farmacia 2.txt
Farmacia 3.txt
Farmacia 4.txt
Farmacia 6.txt
Farmacia 9.txt
Farmacologia 4.txt
Farmacologia 5.txt
Quimica 4.txt
Quimica 8.txt
Quimica 9.txt

Cluster 17:
Enfermagem 7.txt
Enfermagem 8.txt
Raul - 09-04.txt

Cluster 18:
Aquicultura 8.txt
Agronomia 3.txt
Aquicultura 10.txt
Aquicultura 2.txt
Aquicultura 3.txt
Aquicultura 4.txt
Aquicultura 5.txt
Aquicultura 6.txt
Aquicultura 7.txt
Aquicultura 9.txt
Bioquimica 3.txt
Ricardo PS - 14-04.txt

Cluster 19:
Agronomia 8.txt
Agronomia 2.txt
Agronomia 7.txt
Ciencias Humanas 10.txt
Ciencias Humanas 6.txt
Historia 14.txt
Historia 3.txt
Literatura 10.txt
Ricardo S - 14-04.txt

Cluster 20:
Carla - 06-04.txt
Carina - 06-04.txt
Chris - 09-09.txt
Eng Civil 3.txt
Eng Mecanica 9.txt
Guto - 06-04.txt
Mario - 09-04.txt
Ronaldo - 14-04.txt
Vania - 14-04.txt
west - 9-04.txt

Total de clusters: 20
Total de clusters com mais de um arquivo: 20

APÊNDICE D - Resumo do Experimento 2 com Integração ao LExR

Considerando intervalos de 7 anos entre 2009 - 2015.
Período de 2009 - 2015:
Definidos 24 clusters para 158 currículos.

Cluster 1:
- Ricardo PS - 14-04.txt

Cluster 2:
- Farmacologia 7.txt

Cluster 3:
- Farmacologia 2.txt

Cluster 4:
- Aquicultura 4.txt
- Aquicultura 8.txt

Cluster 5:
- Enfermagem 13.txt
- Agronomia 7.txt
- Enfermagem 12.txt
- Enfermagem 14.txt

Cluster 6:
- Aquicultura 7.txt
- Agronomia 3.txt
- Agronomia 4.txt
- Aquicultura 10.txt
- Aquicultura 2.txt
- Aquicultura 3.txt
- Aquicultura 5.txt
- Aquicultura 6.txt
- Aquicultura 9.txt
- Bioquimica 3.txt
- Eng Alimentos 4.txt
- Eng Alimentos 8.txt
- Eng Civil 9.txt

Cluster 7:
- Eng Civil 11.txt
- Eng Civil 12.txt
- Eng Civil 2.txt
- Eng Civil 3.txt
- Eng Civil 5.txt
- Eng Mecanica 1.txt
- Eng Mecanica 3.txt
- Eng Mecanica 4.txt
- Eng Mecanica 5.txt
- Eng Mecanica 9.txt

Cluster 8:
- Eng Alimentos 12.txt
- Eng Alimentos 3.txt
- Eng Alimentos 7.txt
- Eng Alimentos 9.txt
- Eng Civil 4.txt
- Eng Civil 7.txt
- Eng Mecanica 6.txt
- Farmacia 1.txt

Cluster 9:
- Farmacia 8.txt
- Farmacia 5.txt
- Farmacia 6.txt
- Farmacia 7.txt
- Farmacologia 3.txt

Cluster 10:
- Farmacologia 11.txt
- Bioquimica 1.txt

Eng Alimentos 5.txt
Farmacologia 1.txt
Farmacologia 6.txt
Cluster 11:
Farmacologia 5.txt
Farmacia 9.txt
Farmacologia 8.txt
Cluster 12:
Enfermagem 7.txt
Enfermagem 2.txt
Enfermagem 3.txt
Enfermagem 8.txt
Filosofia 10.txt
Cluster 13:
Enfermagem 1.txt
Filosofia 4.txt
Cluster 14:
Fisica 1.txt
Fisica 3.txt
Fisica 4.txt
Cluster 15:
Historia 2.txt
Agronomia 2.txt
Agronomia 5.txt
Ciencias Humanas 1.txt
Ciencias Humanas 10.txt
Ciencias Humanas 2.txt
Ciencias Humanas 8.txt
Filosofia 11.txt
Historia 11.txt
Historia 12.txt
Historia 13.txt
Historia 14.txt
Historia 15.txt
Historia 3.txt
Historia 7.txt
Historia 8.txt
Historia 9.txt
Cluster 16:
Literatura 11.txt
Filosofia 8.txt
Literatura 10.txt
Literatura 2.txt
Literatura 7.txt
Literatura 8.txt
Cluster 17:
Aldo - 06-04.txt
Chris - 09-09.txt
Mario - 09-04.txt
Cluster 18:
Bioquimica 2.txt
Bioquimica 4.txt
Bioquimica 8.txt
Farmacia 12.txt
Farmacologia 12.txt
Quimica 2.txt
Cluster 19:
Quimica 11.txt
Bioquimica 7.txt
Fisica 10.txt
Quimica 1.txt
Quimica 10.txt
Quimica 12.txt
Quimica 13.txt
Quimica 3.txt
Cluster 20:
Quimica 9.txt
Eng Alimentos 10.txt

Eng Alimentos 6.txt
Eng Mecanica 11.txt
Fisica 8.txt
Quimica 5.txt

Cluster 21:
Farmacologia 10.txt
Eng Alimentos 1.txt
Eng Alimentos 11.txt
Eng Alimentos 2.txt
Eng Civil 1.txt
Farmacia 10.txt
Farmacia 2.txt
Farmacia 3.txt
Farmacia 4.txt
Farmacologia 4.txt
Fisica 11.txt
Fisica 5.txt
Fisica 7.txt
Quimica 14.txt
Quimica 4.txt
Quimica 8.txt

Cluster 22:
Enfermagem 9.txt
Agronomia 6.txt
Agronomia 8.txt
Ciencias Humanas 7.txt
Enfermagem 10.txt
Enfermagem 11.txt
Enfermagem 4.txt
Enfermagem 5.txt
Enfermagem 6.txt
Filosofia 6.txt
Literatura 6.txt
Ricardo S - 14-04.txt

Cluster 23:
Ronaldo - 14-04.txt
Carina - 06-04.txt
Ciencias Humanas 6.txt
Elder - 14-04.txt
Filosofia 1.txt
Filosofia 5.txt
Filosofia 9.txt
Historia 4.txt
Lau - 09-04.txt
Literatura 4.txt
Literatura 5.txt
Luiz - 09-04.txt
Mauro - 09-04.txt
Raul - 09-04.txt
Vania - 14-04.txt

Cluster 24:
Carla - 06-04.txt
Ciencias Humanas 5.txt
Eng Mecanica 12.txt
Guto - 06-04.txt
Quimica 7.txt
west - 9-04.txt

Total de clusters: 24
Total de clusters com mais de um arquivo: 21

APÊNDICE E - Artigo

tCALC: Agrupamento de Currículos Lattes por Afinidade de Áreas de Conhecimento Considerando Temporalidade

Jaime Mendes da Silva

Universidade Federal de Santa Catarina
Florianópolis, BR.

jaimemnds@hotmail.com

Abstract. *Works published before have clustered Lattes curricula from science, technology and innovation's field's professionals through the application of data clustering algorithms [1]. The clusters generated by this process have evidenced information about the field they were working in and which of them were on the same field. The current project extends what have been done by analyzing the impact onto quality and performance caused by the consideration of the temporal aspect of the data in the curricula clustering. The inclusion of time in this application comes from the evidence found in the literature that the expertise retrieval applications have benefit from this inclusion [3]. The effort comes from the fact that professionals that worked in some research field in the past might no longer work on the same subject.*

Resumo. *Trabalhos realizados anteriormente, através de algoritmos de clustering de dados, agruparam currículos Lattes de profissionais da área de ciência, tecnologia e inovação [1]. Os grupos gerados por esse processo evidenciavam informações sobre a área de atuação desses profissionais e quais pertencem a uma mesma área. O presente trabalho estende o que foi realizado ao analisar o impacto de qualidade e performance causado pela consideração do fator tempo no processo de agrupamento dos currículos. A inclusão da temporalidade vem da evidência na literatura de que aplicações de busca por competências se beneficiaram da mesma [3]. A aplicação dá-se pelo fato de que profissionais que atuaram em determinada área do conhecimento no passado podem não ser mais atuantes na mesma.*

1. Introdução

Dados são elementos chave de sistemas computacionais [4]. A importância dos computadores nas atividades humanas, somada ao grande avanço tecnológico dos recursos computacionais e sua consequente redução de custos, trouxe como consequência um estado de geração rápida, variada e massiva de dados [4].

Produzir dados não garante por si só que informação seja adquirida a partir deles e que algum conhecimento seja obtido a partir dessa informação. Para preencher essa lacuna, surgiram diversas disciplinas que se propõem a tratar os dados de forma a torná-los úteis para o uso humano em suas diversas aplicações. Entre elas, a Análise de Dados e suas diversas técnicas [7].

Uma importante técnica para o escopo deste trabalho é a Mineração de Dados (do inglês *data mining*). Ela consiste do processo de descoberta de padrões interessantes acerca de uma grande quantidade de dados [2]. Uma dos métodos de Mineração é o agrupamento (*clustering*), que procura particionar os dados evidenciando grupos que concentram todas aquelas amostras que demonstram comportamentos similares em relação a determinadas propriedades [2]. O agrupamento é um dos temas centrais deste trabalho.

A fonte de dados utilizada neste trabalho são currículos da plataforma Lattes, que reúne bases de dados de currículos profissionais da área de pesquisa, desenvolvimento e inovação no Brasil. A informação que será buscada a partir deles são as competências dos autores desses currículos.

Este trabalho parte da obra "CALC: Agrupamento de Perfil Científico por Afinidade de Áreas de Conhecimento Utilizando Currículos Lattes", por Bernardo de Farias Esteves (2015). A introdução do aspecto temporal ao problema que já havia sido abordado no projeto CALC dá-se com base em linhas recentes de pesquisa que propõem que o tempo é um fator com alto teor de relevância para avaliar a perícia da qual um profissional dispõe em relação a uma área [3].

2. Descrição do Processo

As atividades deste trabalho seguem as etapas definidas pelo processo de KDD (*Knowledge Discovery in Databases* - Descoberta de Conhecimento em Bancos de Dados). Esse processo foi desenvolvido para servir de ferramenta a auxiliar na extração de informações úteis do grande volume de dados digitais produzidos atualmente e, principalmente, de dados que, se considerados individualmente, possuam baixo teor informativo [6].

O KDD é composto por 5 fases que se fazem presentes no procedimento deste trabalho (seleção, pré-processamento, transformação, mineração de dados, interpretação/avaliação) que, embora propostas nessa sequência, são organizadas de maneira diferente conforme as necessidades do projeto, o que não afeta o resultado.

2.1. Seleção

A primeira etapa que o KDD sugere é a seleção, dentre todo o conjunto de dados disponível, daqueles que são relevantes para a produção do conhecimento. Essa etapa também pode considerar a seleção de dados adicionais de fontes externas [4].

O primeiro aspecto dessa etapa, em relação ao procedimento deste projeto, foi a definição da informação que se pretende extrair do conjunto de dados sobre os quais será trabalhado, ou seja, “o que queremos saber sobre os dados?”.

A resposta a essa pergunta herda seus fundamentos do projeto executado anteriormente por Esteves (2015), onde a resposta era “a área de atuação de um dado profissional” ou “quais profissionais atuam na mesma área”. Mas no caso atual a resposta se expande: queremos saber dos dados quais profissionais atuaram (ou têm competência) na mesma área em um mesmo momento do tempo.

A partir da diferença dessas propostas, foi decidido que o tCALC deveria tratar a etapa de Seleção de maneira diferente, buscando campos que evidenciassem as informações sobre área de atuação e período de atuação em paralelo. Para isso, decidiu-se utiliza a seção referente às publicações do profissional, que trazia essas informações de maneira consistente, mais especificamente os campos ‘TITULO-DO-ARTIGO’ e ‘ANO-DO-ARTIGO’.

Como já se esperava, essa escolha mostrou fragilidades desde os primeiros experimentos que seguiram sua implementação. A pequena quantidade de dados disponível nesses dois atributos impactou diretamente na qualidade do agrupamento. Para melhorar isso,

a base de dados *Lattes Expertise Retrieval* (LExR) [8] foi escolhida para a obtenção de dados adicionais acerca dos artigos considerados na entrada.

2.2. Pré-Processamento

É indicado que haja um tratamento adequado dos dados, sejam eles estruturados ou não, antes do processo de mineração [5]. No caso do projeto tCALC as atividades de pré-processamento se limitam à ação conhecida como remoção de *stopwords*. Essa ação é bastante comum em mineração de dados textuais e consiste na remoção daquelas estruturas das linguagens naturais que conectam as ideias dentro de uma frase mas que, por si só, não apresentam muito valor semântico para a mineração. Exemplos comuns desses conectivos são as preposições, artigos e pronomes [10].

2.3. Transformação

Após a fase de pré-processamento garante-se que os dados tenham adquirido uma condição consistente, mas não garante que eles estejam prontos para a mineração. A última ação a ser tomada antes de aplicar os algoritmos de mineração deve ser a transformação dos dados de seu formato original em um que os algoritmos de mineração aceitem como entrada, o que costuma variar entre diferentes implementações [5].

O tCALC executa a transformação através da exportação dos dados referentes ao campos ‘TITULO-DO-ARTIGO’ e ‘ANO-DO-ARTIGO’ para arquivos em disco. Após isso, é feita uma consulta no banco de dados referente ao LExR para cada artigo, incluindo no arquivo de texto as palavras-chave e áreas obtidas da base de dados.

2.4. Mineração

Esta é, provavelmente, a etapa mais sofisticada do ponto de vista computacional no KDD. Segundo Fayyad et al (1996), a “mineração de dados é a aplicação de algoritmos específicos para extrair padrões a partir de dados”.

A mineração pode ser feita de diversas formas, através de diversas técnicas e ainda, cada técnica pode ser implementada por algoritmos diversos e por vezes bem distintos [4]. A técnica abordada no procedimento deste trabalho foi idealizada por Esteves (2015) e consiste no agrupamento de dados usando os algoritmos *BestStar* e *K-medoids* implementados pelo URSA, um *framework* que oferece algoritmos para cálculo de similaridade e agrupamento de dados e que permite a aplicação dessas funcionalidades sobre diversos tipos de dados [1].

O tCALC, a exemplo do CALC, utiliza o algoritmo *BestStar* inicialmente para estimar uma quantidade ótima de *clusters* a serem gerados para o conjunto de dados de entrada. Essa etapa é uma preparação para a execução do *K-medoids* já que esse segundo algoritmo espera que seja fornecida como entrada a quantidade de *clusters* que serão gerados para os dados, enquanto o *BestStar* não precisa dessa informação [1].

2.5. Análise/Interpretação

A última fase do processo é mais dependente de intervenção humana que as demais. As saídas dos algoritmos da fase anterior são tratadas de forma a evidenciar as informações obtidas através de gráficos ou qualquer outro tipo de representação que seja mais simpática à análise humana [4].

Após a disponibilização visual desses resultados, é necessário que o analista os avalie e compare na tentativa de obter informações evidenciadas pelos padrões formados ou,

mediante a devida constatação, decida pela alteração do processo para uma nova tentativa de KDD [4].

Neste projeto, a aplicação gera como saída do agrupamento uma hierarquia de diretórios no sistema de arquivos que revelam os *clusters* formados. Essa árvore de diretórios se organiza com uma raiz chamada *Clusters*, dentro da qual existem pastas referentes a cada período de tempo considerado no agrupamento e denominadas com base no primeiro ano do intervalo considerado (por exemplo, a pasta referente ao triênio 2010-2012 é denominada 2010). Dentro da pasta de cada ano estão diretórios numerados referentes aos *clusters* gerados nos quais estão os currículos Lattes atribuídos a esses grupos.

As análises deste projeto foram feitas com base na saída descrita acima a fim de gerar deduções sobre a qualidade dos *clusters* gerados e da possibilidade de aprimorar o processo para melhorar os resultados.

3. Experimentos

A amostra de currículos utilizada possui 206 arquivos contendo currículos Lattes no formato XML de pesquisadores de programas de pós-graduação da UFSC das áreas: agronomia, aquicultura, bioquímica, ciência da computação, ciências humanas, enfermagem, engenharia de alimentos, engenharia civil, engenharia mecânica, farmácia, farmacologia, filosofia, física, história, literatura e química. Essa amostra foi a mesma utilizada na implementação do CALC e foi escolhida pelo autor daquele projeto de forma a possuir elementos de áreas distintas mas também de áreas com alguma semelhança para analisar efeitos que mudanças nos parâmetros causariam aos resultados do agrupamento [1].

Os currículos da amostra são identificados pelos nomes das áreas aos quais pertencem, seguidos de números para diferenciar uns dos outros (por exemplo, Agronomia 3 ou Aquicultura 1). Alguns outros currículos, pertencentes a profissionais de Ciência da Computação, são denominados com os nomes dos profissionais que retratam. Essas estratégias foram tomadas para permitir que, na etapa de Análise, seja possível distinguir quais *clusters* possuem currículos semelhantes de fato [1].

3.1. Experimentos sem LExR

Nesta fase os experimentos foram elaborados buscando avaliar a qualidade do agrupamento ainda sem a intergração com a base de dados externa e, para isso, utilizou-se o lote de amostras com 206 currículos.

O experimento realizado, considerou triênios para agrupamento. Ou seja, fez um agrupamento para cada triênio contido no período no qual a amostra trouxe artigos publicados.

3.2. Experimentos com LExR

Os experimentos desta etapa foram bastante semelhantes aos anteriores, visto que a única mudança – embora substancial – foi a adição dos dados da base de dados externa LExR no processo. Com os dados adicionais de palavras-chave e áreas dos artigos do Lattes, foi possível aumentar a expressividade dos dados de entrada.

A próxima subseção se preocupa em avaliar a qualidade alcançada nesses experimentos.

3.3. Análise de Resultados

A análise dos resultados deste projeto dá-se de forma puramente argumentativa com base nos experimentos apresentados.

Os currículos fornecidos como entrada da aplicação foram nomeados com as áreas dos profissionais que eles representam com o intuito de facilitar a observação de *clusters* gerados com sucesso ou não. A argumentação presente nestas análises parte da definição que as seguintes áreas presentes no conjunto de entrada são mutuamente próximas no ponto de vista da engenharia de conhecimento:

- Química, Bioquímica, Farmácia e Farmacologia;
- Agronomia e Aquicultura;
- Ciências Humanas, Filosofia, História e Literatura;
- Física e Engenharia Civil;
- Física e Engenharia Mecânica;
- Física e Química;
- Química, Física e Engenharia de Alimentos.

Outras relações não listadas podem ser verdadeiras mas não foram consideradas nas análises a seguir.

As análises foram feitas para todos os *clusters* gerados nos experimentos e levaram em consideração a taxa de agrupamentos bem-sucedidos – ou seja, aqueles que colocam em um mesmo grupo currículos de áreas próximas conforme a lista acima – e, para quando ocorrer, a quantidade de subgrupos formados no *cluster*. Os resultados com ou sem a integração com LExR são comparados ao final de cada experimento.

Os dados abaixo dizem respeito ao triênio entre 2011 a 2013. E foram obtidos pela implementação sem uso do LExR:

- Quantidade total de currículos: 165;
- Quantidade total de *clusters*: 11;
- Quantidade de currículos agrupados corretamente: 127;
- Quantidade de currículos agrupados incorretamente: 38;
- Quantidade de subgrupos: 10 divididos entre 4 grupos;
- Taxa de sucessos: 77%;
- Taxa de fracassos: 23%.

Os dados obtidos pela execução integrada ao LExR são:

- Quantidade total de currículos: 165;
- Quantidade total de *clusters*: 14;
- Quantidade de currículos agrupados corretamente: 142;
- Quantidade de currículos agrupados incorretamente: 23;
- Quantidade de subgrupos: 16 divididos entre 7 grupos;
- Taxa de sucessos: 86%;
- Taxa de fracassos: 14%.

Portanto, com base nas variáveis enumeradas acima, pode-se concluir que há uma melhora com o uso do LExR no experimento. A taxa de sucessos de 86% é considerada satisfatória visto as limitações apresentadas pela base de dados tais como a escassez de termos e campos com informação temporal. A grande quantidade de subgrupos evidencia que, se fosse assumido um número maior de grupos, esses subgrupos se manifestariam como grupos por si só, o que indica que a quantidade ótima de grupos deve ser maior que a escolhida pelo algoritmo. Isso pode abrir espaço para otimizações no processo.

4. Considerações Finais

A aplicação produzida juntamente com esta obra se preocupou, a exemplo de sua antecessora, em oferecer a solução a uma demanda inerente à proposta da plataforma Lattes, o que literatura chama de *expertise retrieval* ou recuperação de especialidades em uma tradução livre. Além disso, o projeto em si buscou apresentar a importância da mineração de dados, da descoberta de conhecimento e da análise de competências.

Até esses requisitos apresentados anteriormente, o projeto tCALC apenas expandiu o que já havia sido abordado no CALC. A diferença fundamental e que permitiu que um novo escopo fosse trabalhado – o aspecto temporal – foi abordada com protagonismo compatível com a relevância dela para o problema apresentado. Ao optar-se por levar em consideração o tempo no *expertise retrieval*, a proposta muda radicalmente tal como os resultados atingidos. E isso foi o que este projeto se preocupou em deixar claro através de seus capítulos de experimentos e análises.

Ainda há muito o que ser produzido pela comunidade científica nas áreas às quais este trabalho pertence. Ainda assim, a expectativa é de que, de alguma forma, esta obra tenha contribuído, seja para gerar conhecimento acerca do assunto ou motivação para que outros projetos avancem cada vez mais o estado da arte.

Referências Bibliográficas

[1] B. F. ESTEVES. *CALC: Agrupamento de Perfil Científico por Afinidade de Áreas de Conhecimento Utilizando Currículos Lattes*. Florianópolis: Universidade Federal de Santa Catarina. 2015.

[2] M.J. ZAKI, and W. MEIRA. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. 1st ed. New York: Cambridge UP, 2014.

[3] Y. LI, and J. TANG. *Expertise Search in a Time-varying Social Network*. Beijing: Tsinghua University, 2008.

[4] U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH, and R. UTHURUSAMY. *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA 1-34. 1996.

[5] S. GARCÍA, J. LUENGO, and F. HERRERA. *Data Preprocessing in Data Mining.* In *Intelligent Systems Reference Library vol. 72.* New York: Springer, 2015.

[6] A. ADHIKARI, and J. ADHIKARI. *Advances in Knowledge Discovery in Databases.* In *Intelligent Systems Reference Library vol. 79.* New York: Springer, 2015.

[7] C.R. SHALIZI. *Advanced Data Analysis from an Elementary Point of View*. New York: Cambridge UP. 2016.

[8] V. MAGARAVITE, R.L.T. SANTOS, I.S. RIBEIRO, M.A. GONÇALVES, A.H.F. LAENDER. *The LExR Collection for Expertise Retrieval in Academia.* Department of Computer Science, Universidade Federal de Minas Gerais. Belo Horizonte, MG, Brazil.