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]. Um 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, possuem 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 utilizar 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.