Mineração em dados educacionais para avaliar evasão de alunos da área de tecnologia da informação

Kelvyn Yago da Silva Zanato[1], **Thiago Meirelles Ventura**[1]

[1] Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Campus Cuiabá – MT – Brazil

kelvynzanatoo@gmail.com, thiago@ic.ufmt.br

Abstract. *The evasion of students at the higher level is a serious problem. This work introduces a method based on graph-oriented Database to perform identification of students with high evasion risk. Our approach uses exclusively the students' academic history, something easy to get to applied the proposed method. It was calculated the similarity between current students and evaded students from previous classes. The results showed that is possible to accurately identify 73% the final situation of the student. The proposed method may indicate the taking of individual actions directed at students at risk, as well as the planning of future actions, to reduce the amount of evaded students.*

Resumo. *A evasão de estudantes no nível superior é um sério problema. Neste trabalho, é apresentado uma metodologia baseada em banco de dados orientado a grafos com o intuito de identificar estudantes com potencial de desistência. Foi utilizado exclusivamente o histórico escolar dos alunos, algo facilmente obtido para a aplicação da metodologia proposta. Cálculos de similaridade entre alunos atuais com alunos evadidos de turmas anteriores foram realizados. Os resultados mostraram que é possível identificar com precisão média de 73% a situação final do aluno. A metodologia proposta pode auxiliar a tomada de ações individuais direcionadas a alunos em risco, bem como o planejamento de ações futuras, a fim de diminuir a evasão dos estudantes.*

Palavras-chave*: Banco de Dados Não Convencionais, Banco de dados orientado a grafos, Neo4j, Educação, Evasão.*

1. Introdução

A evasão de estudantes nas universidades brasileiras é um problema que atinge praticamente todos os cursos de nível superior, tendo como algumas consequências a formação de profissionais abaixo da capacidade desejada, frustração dos alunos que não conseguem concluir sua graduação e significativo desperdício de recursos [Hipólito 2011].

A evasão está relacionada a vários fatores, que podem ser divididos entre internos e externos. Os fatores internos são ligados ao curso, e podem ser classificados em: infraestrutura, corpo docente e a assistência sócioeducacional. Os fatores externos relacionam-se ao aluno, tais como: vocação, aspectos socioeconômicos e problemas pessoais [Paredes 1994].

É preciso fazer esforços que evitem ou diminua a ocorrência de evasão e também de retenção em cursos superiores, em especial as instituições públicas, como forma

de evitar a perda de recursos públicos. Esses esforços devem passar por processos de observação e compreensão dos aspectos, tanto gerais quanto específicos, levando em consideração as peculiaridades de cada curso e outros contextos, como os sociais, de ensino e administrativo de alunos e profissionais envolvidos na educação superior [Lobo 2012, Manhães et al. 2011].

Com o intuito de elucidar parte das situações que configuram essa problemática, é importante a construção de instrumentos para acompanhar a trajetória dos estudantes. Desta forma, pode ser possível identificar antecipadamente aqueles com grande potencial de desistência ou desligamento. Tais instrumentos têm importância tanto na tomada de ações individuais (aconselhamento e orientação do aluno) como também na previsão de retenções futuras e planejamento do oferecimento de turmas extras.

Este trabalho tem como objetivo analisar dados de alunos para identificar padrões e prever relações. Dados como reprovação em disciplinas, curso, tipo de ingresso, quantidade de alunos por turma, entre outros podem servir como características para definir alunos que desistirão do curso. No entanto, ao construir consultas para obter estas relações em um banco de dados relacional pode-se exigir diversos junções e subconsultas, podendo tornar a consulta de extrema complexidade e com baixa performance. Portanto, foi definido neste trabalho o uso de um banco de dados orientado a grafos, no qual é possível conseguir os resultados desejados.

2. Trabalhos Correlatos

O estudo do problema de evasão escolar possibilita identificar a sua relação com uma demanda importante na sociedade, onde tanto universidades públicas como universidades particulares apresentam índices considerados altos. Diversos trabalhos vêm sendo realizados com o objetivo de identificação de tendências de evasão [Rigo et al. 2012]. Em [Manhães et al. 2014] é apresentado uma arquitetura que utiliza técnicas de mineração de dados para monitorar o progresso acadêmico dos estudantes e prever a aprovação dos mesmos nas disciplinas. Os alunos que não serão aprovados possuem um risco maior de evasão. Algoritmos de classificação como *Naive Bayes* foram utilizados para obtenção dos resultados.

Em [Santos et al. 2015] foi realizado um estudo utilizando a análise do motivo da evasão e a aplicação de um questionário para alunos evadidos. Neste questionário havia diversas questões referentes a aptidões, anseios, repercussão da evasão no aluno e possíveis intervenções no curso. Para o curso identificou-se como principais causas de evasão a insuficiência da infraestrutura de apoio ao ensino de graduação e a exigência de dedicação exclusiva ao curso. Já em [Fernandes and Junior 2016] foi feito uma estudo analisando documentos produzidos pelos professores com o intuito de comprovar a relação entre evasão e reprovação de disciplinas como lógica e programação.

Como pode ser visto, o problema de evasão está presente em várias instituições, e os pesquisadores estão utilizando diversos métodos para solucionar o problema. Os trabalhos citados auxiliam a coordenação de curso no contexto de evasão. Entretanto, ainda é necessário um método para, de forma mais específica, identificar os alunos que tendem a evadir. Este trabalho utiliza um banco de dados orientado a grafos para identificar alunos com risco de evasão fazendo uma análise das relações entre os dados do estudante.

3. Materiais e Métodos

Para o desenvolvimento deste trabalho foi necessário realizar a definição da tecnologia, obtenção dos dados brutos, organização dos dados, importação dos dados, criação de consultas orientadas a grafos e análise dos resultados. A seguir são descritos com mais detalhes cada um desses itens.

3.1. Tecnologias Utilizadas

O banco de dados selecionado foi o Neo4j. Este banco tem uma estrutura diferenciada dos bancos de dados relacionais no que se refere às relações dos dados. Ele guarda seus dados em forma de grafos, uma forma elegante de armazenar e relacionar qualquer tipo de informação por mais abstrata que ela seja de uma forma acessível através das relações entre os nós. A tecnologia de banco de dados em grafos é uma ferramenta eficaz para a modelagem de dados quando o foco no relacionamento entre as entidades é uma força motriz na concepção dos dados [Miller 2013].

A vantagem de utilização do modelo baseado em grafos fica bastante clara quando consultas complexas são exigidas pelo usuário. Comparado ao modelo relacional, que para estas situações pode ser muito custoso, o modelo orientado a grafos tem um ganho de performance, permitindo um melhor desempenho das aplicações [Lóscio et al. 2011]. Para tanto, é necessário uma linguagem específica para poder interagir com o banco. De acordo com [Eduardo 2015], *Cypher* é uma linguagem para executar as *queries* no Neo4j. É parecida com a linguagem SQL, porém com propriedades únicas referentes ao banco de dados grafo:

- Em banco de dados relacionais existem tabelas e registros, já no *Cypher* as tabelas são identificadas como *labels* e registros como *nodes*;
- As relações entre as *labels* se chamam *relationship* que podem ou não ter propriedades;
- Uma tabela em banco de dados relacionais existem campos que representam valores dos registros, ex: nome, idade, sexo. No *Cypher* existe um node no formato *key-value* ou seja, *object*, ex: {nome:'valor', idade:valor}.

Como exemplo, pode ser feito uma instrução que lista os funcionários que trabalham no departamento de TI. Em SQL, a consulta seria:

```
SELECT nome FROM Pessoa
LEFT JOIN Pessoa_Departamento
   ON Pessoa.Id = Pessoa_Departamento.PessoaId
LEFT JOIN Departamento
   ON Departamento.Id = Pessoa_Departamento.DepartamentoId
WHERE Departamento.nome = "Tecnologia da Informacao"
```

Já no Cypher, a mesma instrução seria:

```
MATCH (p:Pessoa)-[:TRABALHA]->(d:Departamento)
WHERE d.nome = "Tecnologia da Informacao"
RETURN p.nome
```

Nesse exemplo, a consulta *Cypher* é mais simples e clara comparada à instrução. Não só a consulta *Cypher* será mais rápida para criar e executar, mas também reduz as chances de resultados inesperados [Michael Hunger 2016].

3.2. Obtenção e Descrição dos Dados Brutos

A base de dados utilizada neste trabalho foi fornecida pelo Profº. Jivago Medeiros Ribeiro[1] do IC da UFMT elaborada a partir de dados extraídos do Sistema de Informações de Gestão Acadêmica (SIGA) da universidade para um arquivo CSV.

A Tabela 1 contém um exemplo de como os dados estão distribuidos no arquivo.

Tabela 1. Distribuição dos dados no arquivo CSV

RGA	TIPO_INGRESSO	SEMESTRE	SEM_COUNT	NOME_DISCIPLINA	MEDIA	SITUACAO	EVADIDO
200000000001	1	2000/2	2	VETORES_E_GEOMETRIA_ANALITICA	8.25	AP	SIM
200100000002	3	2001/1	1	FUNDAMENTOS_DA_COMPUTACAO	6.00	AP	SIM
200200000003	2	2002/2	2	ALGEBRA_LINEAR	7.00	AP	SIM
200300000004	5	2003/1	1	FILOSOFIA_DA_CIENCIA	9.50	AP	SIM
201200000005	6	2013/2	4	LINGUAGEM_DE_PROGRAMACAO_II	0.00	AE	SIM

Foram obtidos um total de 31.332 registros de históricos escolares dos alunos e ex-alunos dos cursos de Ciências da Computação e Sistemas de Informação do Instituto de Computação da UFMT no período entre 2000 e 2015.

3.3. Organização e Importação dos Dados

Com os dados obtidos, a modelagem em grafo foi realizada, relacionado os alunos às suas disciplinas. Nesse tipo de banco, as relações também contém informações. Para este trabalho, a relação de ”cursar”possui informações sobre período, nota e situação que esse relacionamento aconteceu. Além disso, foi feito um tratamento nos dados, transformando a coluna SEMESTRE em ANO e PERIODO (separando-os através da barra) e removendo os caracteres especiais da coluna NOME_DISCIPLINA, substituindo-os por espaços. Após estes passos os dados foram importados utilizando a própria ferramenta disponibiliza pelo Neo4j. Um exemplo das relações de um aluno é mostrado na Figura 1.

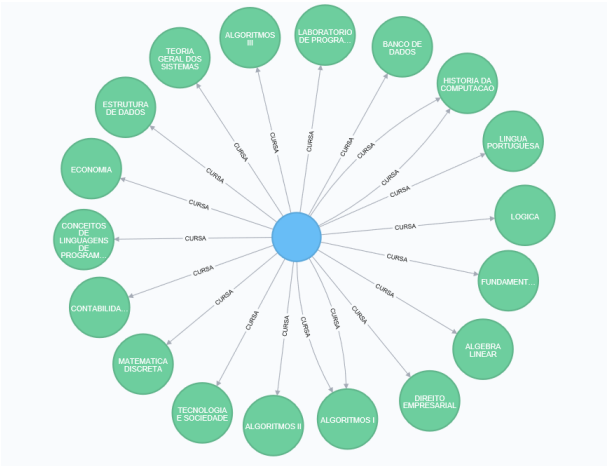

Figura 1. Representação em grafo de disciplinas cursadas por um aluno.

4. Execução dos Testes e Análise dos Resultados

Nesta subseção são apresentadas as consultas criadas na linguagem *Cypher* e seus resultados com o objetivo de analisar questões que envolvem a evasão. Uma das questões

[1] jivago@ic.ufmt.br

a serem discutidas é se alguma disciplina pode ser o motivo de evasão. Para isso, uma consulta foi criada envolvendo a quantidade de reprovações por disciplina.

```
MATCH (a:Aluno)-[c]->(d:Disciplina)
WHERE a.evadiu = TRUE
AND c.situacao IN ["RM","RMF","RF"]
RETURN (d.nome), count (c) AS Total
ORDER BY Total DESC
```

O código faz a consulta de todos os alunos evadidos que possuem relação com disciplina e que foram reprovados (seja por média e/ou falta). Os resultados são apresentados na Tabela 2.

Tabela 2. Disciplinas com maiores índices de reprovações

DISCIPLINA	TOTAL_ALUNOS
ALGORITMOS I	362
LOGICA	308
ALGEBRA LINEAR	264
CALCULO I	233
FUNDAMENTOS DA COMPUTACAO	231

De acordo com a Tabela 2, os maiores índices estão relacionados as disciplinas iniciais (primeiro e segundo semestre) dos cursos, sendo que quatro das disciplinas fazem parte da matriz curricular de ambos os cursos (Algoritmos I, Lógica, Álgebra Linear e Fundamentos da Computação). Apenas Cálculo I faz parte exclusivamente da matriz do curso de Ciências da Computação.

Sabendo-se que a disciplina de Algoritmos I tem o maior índice de reprovação dos alunos evadidos, pode ser criado uma consulta para analisar quantos alunos evadidos reprovaram nesta disciplina.

```
MATCH (a:Aluno)-[c]->(d:Disciplina{nome:'ALGORITMOS I'})
WHERE c.situacao IN ["RM","RMF","RF"]
WITH COUNT(DISTINCT a) AS TotalAlunosRep
MATCH (a:Aluno)-[c]->(d:Disciplina{nome:'ALGORITMOS I'})
WHERE c.situacao IN ["RM","RMF","RF"]
AND a.evadiu = TRUE
RETURN TotalAlunosRep,
  COUNT(DISTINCT a) AS QtdAlunosEvadidos
```

O código retorna o total de alunos distintos que foram reprovados na disciplina de Algoritmos I e destes alunos, quantos foram evadidos. Como resultado, teve-se que 388 alunos foram reprovados, sendo 247 evadidos. Portanto, 63.6% dos alunos que reprovaram em Algoritmos I evadiram. Os índices demonstram valores altos devido a obrigatoriedade dos alunos cursarem a disciplina assim que ingressam nos cursos, já que a mesma faz parte do primeiro semestre da matriz curricular de ambos cursos.

Uma relação ligada com semestre a ser analisada é a quantidade de reprovações que os alunos recebem no mesmo semestre antes de evadir. Com as mesmas condições utilizadas pela consulta anterior e adicionando a condição para retornar o último semestre cursado pelos alunos, foram obtidos os resultados desejados e apresentados na Figura 2.

```
MATCH (a)-[c]->(d)
```

```
WHERE c.situacao IN ["RM","RMF","RF"]
AND a.evadiu = TRUE
WITH a, MAX(c.semestre) AS sem
MATCH (a)-[c]->(d)
WHERE c.semestre = sem
AND c.situacao IN ["RM","RMF","RF"]
RETURN a.rga as Aluno, sem, COUNT(d.nome) AS QtdDiscipRep
ORDER BY Aluno, sem
```

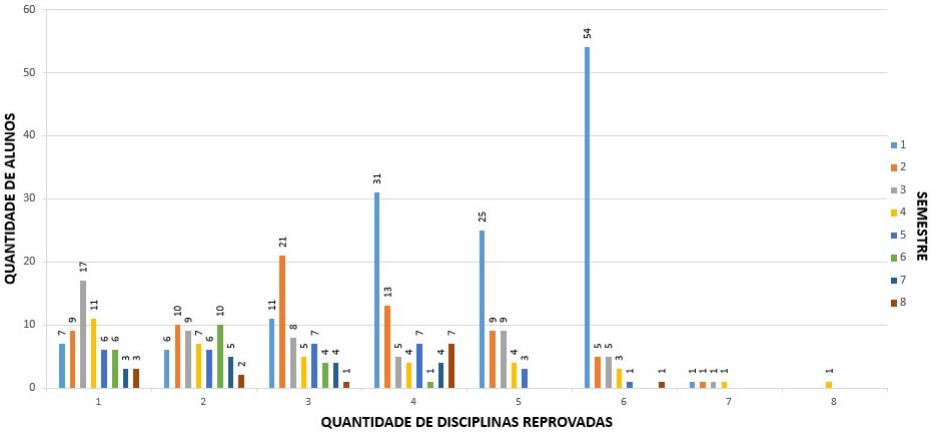

Figura 2. Quantidade de disciplinas reprovadas pelos alunos por semestre.

Analisando a Figura 2, obteve-se um total de 374 alunos evadidos com reprovações em seus históricos. Os maiores índices ocorrem quando os alunos reprovam de 4 ou mais disciplinas no primeiro semestre (110 alunos), representando aproximadamente 30% do total. E analisando apenas o primeiro semestre, dos 135 alunos evadidos que tiveram reprovações neste semestre, 81.5% possuem 4 ou mais reprovações.

Outra questão de extrema importância a ser analisada é a quantidade de reprovações em sequência recebidas pelo alunos antes de evadirem, ou seja, o número de semestres em que ocorreram reprovações. A consulta a seguir recupera esta informação e o respectivo resultado é apresentado na Figura 3.

```
MATCH (a)-[c]->(d)
WHERE c.situacao IN ["RM","RMF","RF"]
AND a.evadiu = TRUE
RETURN DISTINCT a.rga AS Aluno, a.curso AS Curso,
  COUNT(DISTINCT (c.semestre)) AS QtdSem
ORDER BY Aluno
```

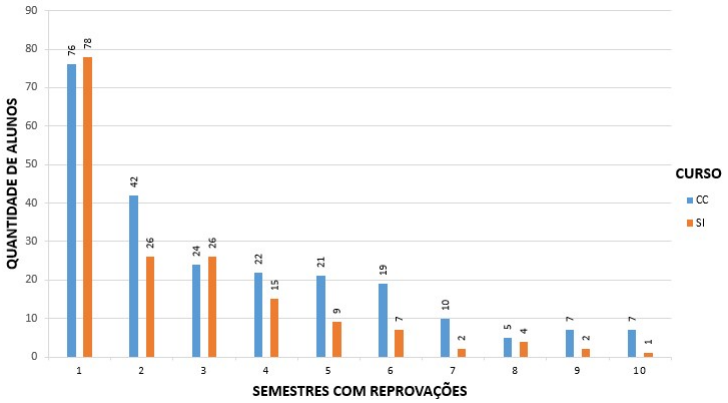

Figura 3. Reprovações em sequência de alunos por curso.

No total foram obtidos 403 alunos evadidos que tiveram reprovações em seus registros históricos, sendo 233 do curso de Ciências da Computação (CC) e 170 do curso de Sistemas de Informação (SI). Portanto, o maior índice de reprovações está entre os três primeiros semestres dos cursos, representando 67.5% do total de alunos evadidos, índice diretamente ligado ao resultado anterior, pois impacta com a quantidade de disciplinas reprovadas no semestre, ou seja, se o aluno obter muitas reprovações, a probabilidade deste aluno evadir é alta.

Juntamente com a relação anterior de sequência de reprovações, pode-se analisar a média de reprovações de uma mesma disciplina entre alunos evadidos.

```
MATCH (a)-[c]->(d)
WHERE c.situacao IN ["RM","RMF","RF"]
AND a.evadiu = TRUE
RETURN a.rga AS Aluno, COUNT(d.nome) AS QtdRep,
  d.nome AS NomeDisc
ORDER BY Aluno, NomeDisc
```

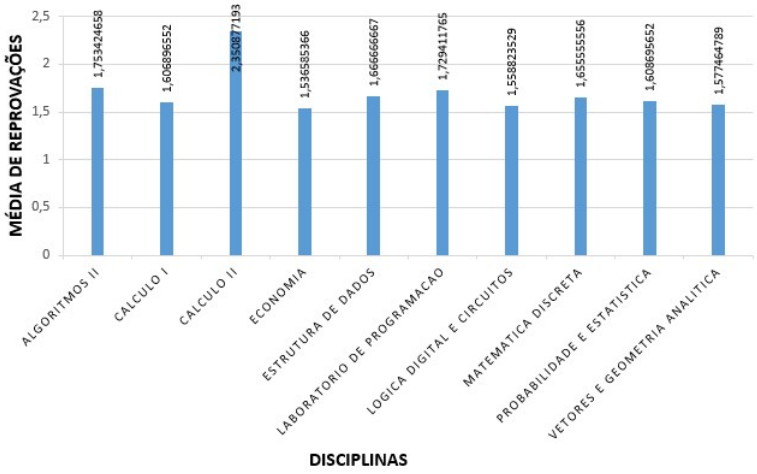

Figura 4. Disciplinas com maiores médias de reprovações de alunos evadidos.

Para obter os resultados da Figura 4, foi realizado um filtro removendo as disicplinas que possuem poucos menos de dez alunos que foram reprovados e também as

disciplinas que não fazem parte da matriz curricular atual dos cursos. De acordo com a Figura 4, pode ser observado uma maior probabilidade dos alunos evadidos reprovam em média duas vezes nas disciplinas de Cálculo I e II, além de pelo menos uma vez nas disciplinas da área de exatas e programação.

Por fim, podemos fazer a previsão de um aluno se ele pode evadir ou não no próximo semestre, de acordo com a similaridade entre alunos evadidos anteriormente. Para isso, a similaridade foi definida como o menor valor absoluto da diferença entre as médias do aluno analisado com alunos das turmas anteriores.

```
MATCH (a)-[r1]->(d)
WHERE toInt(left(toString(a.rga), 4)) = 2013
AND a.curso = 316
AND r1.semestre <= 4
WITH a, SUM(ABS(r1.nota)) / COUNT(r1) as media,
  SUM(ABS(r1.nota)) as soma, COUNT(r1) as qtd
MATCH (a2)-[r2]->(d2)
WHERE toInt(left(toString(a.rga), 4)) >
  toInt(left(toString(a2.rga), 4))
AND a <> a2
AND a.curso = a2.curso
AND r2.semestre <= 4
WITH a, media, soma, qtd, a2,
  SUM(ABS(r2.nota)) / COUNT(r2) as media2,
  SUM(ABS(r2.nota)) as soma2, COUNT(r2) as qtd2,
  COLLECT(r2.semestre) as sem2
MATCH (a3)-[r3]->(d3)
WHERE 4 in sem2
RETURN DISTINCT a.rga, a2.rga, media, media2,
  a.evadiu, a2.evadiu, ABS(media-media2) as similaridade
ORDER BY similaridade, a.rga, a2.rga
```

Neste exemplo, foi utilizado como parâmetro o 4º semestre dos alunos e a turma de 2013 do curso de Sistemas de Informação, devido ao curso possuir poucos registros. A instrução faz a busca destes alunos, trazendo todos os registros históricos até o 4º semestre e, em seguida, calcula a média das notas obtidas com a quantidade de disciplinas cursadas. São obtidos então os alunos de turmas anteriores e do mesmo curso que os alunos da turma de 2013, retornando as notas dos 4 primeiros semestres e também calculando a média de cada aluno. Por fim, finaliza a instrução com um filtro onde verifica os alunos que realmente cursaram 4 semestres (pois no retorno anterior existem alunos que podem ter cursado apenas 1, 2, ou 3 semestres) e é feito o cálculo de similaridade.

Como resultado, foi obtido um total de 4.480 comparações entre os alunos da turma de 2013 com os alunos de turmas anteriores. Dessas comparações, foi filtrado apenas a maior similaridade de cada aluno da turma de 2013 para realizar a análise, ou seja, cada aluno analisado foi comparado com o aluno de turmas anteriores que possuem o resultado de similaridade próximo ao deste aluno, obtendo-se no final 40 resultados, representados na Tabela 3.

Tabela 3. Análise de acerto do cálculo de similaridade entre alunos.

Situação	Total Alunos	Acertos Similaridade	%Acertos
Evadidos	22	15	68,2
Não Evadidos	18	14	77,8

De acordo com a Tabela 3, dos 40 alunos da turma de 2013, 22 realmente evadiram e 18 não evadiram (pelo menos até o período de 2015), sendo que dos 22 alunos evadidos, obteve-se 15 acertos. A mesma análise equivale aos alunos não evadidos, no qual dos 18 que ainda não evadiram, obtiveram similaridade com 14 alunos das turmas anteriores que não evadiram. Desta forma, é possível obter uma estimativa de alunos com risco de evasão após um determinado período de vida acadêmica, fornecendo uma informação importante aos coordenadores de cursos.

5. Conclusão

Este trabalho teve como objetivo identificar causas e fazer previsões de evasão nos cursos de Ciências da Computação e Sistemas de Informação da UFMT, analisando dados de turmas anteriores utilizando banco de dados orientado a grafos, Neo4j. A identificação dos alunos que apresentam risco de evasão por meio do uso de grafos mostrou-se possível. Os experimentos retornaram dados com precisão média de 73%. Pode ser incluso como taxa de erro da análise os alunos que obtiveram aproveitamento de disciplinas, pois no sistema acadêmico algumas dessas disciplinas são lançadas com nota zero. Além disso, pode ser incluído como influência nos resultados os alunos considerados *outliers* que evadem do curso mesmo com alto rendimento acadêmico e alunos que concluem o curso com o rendimento abaixo da média.

O trabalho indica que é possível fazer a previsão de alunos com risco de evasão com base na função de similaridade com alunos de turmas anteriores de acordo com o respectivo curso. Com os resultados foi possível definir também relações entre os cursos e suas disciplinas, podendo auxiliar os coordenadores e a própria instituição na tomada de decisões para diminuir o número de evasões.

Referências

Eduardo, N. (2015). Bem vindo ao neo4j. `http://nicholasess.com.br/neo4j-2/bem-vindo-ao-neo4j/`. Acesso em: 22/04/2017.

Fernandes, V. d. S. and Junior, V. F. (2016). Evasão e reprovação nas disciplinas de lógica e programação: Informações preliminares no campus sombrio, do instituto federal catarinense. In *Anais do 5º Simpósio de Integração Científica e Tecnológica do Sul Catarinense*, Araranguá-SC.

Hipólito, O. (2011). O gargalo do ensino superior brasileiro: Depoimento. `https://www.cartacapital.com.br/sociedade/o-gargalo-do-ensino-superior-brasileiro`. Entrevista concedida a Fernando Vives. Acesso em: 17/04/2017.

Lobo, M. B. d. C. M. (2012). Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, (25).

Lóscio, B. F., OLIVEIRA, H. R. d., and PONTES, J. C. d. S. (2011). Nosql no desenvolvimento de aplicações web colaborativas. In *VIII Simpósio Brasileiro de Sistemas Colaborativos*, volume 10, page 11.

Manhães, L. M. B., da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., and Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 1.

Manhães, L. M. B., da Cruz, S. M. S., and Zimbrão, G. (2014). Wave: an architecture for predicting dropout in undergraduate courses using edm. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 243–247. ACM.

Michael Hunger, Ryan Boyd, W. L. (2016). Rdbms & graphs: Sql vs. cypher query languages. `https://neo4j.com/blog/sql-vs-cypher-query-languages/`. Acesso em: 11/05/2017.

Miller, J. J. (2013). Graph database applications and concepts with neo4j. In *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*, volume 2324, page 36.

Paredes, A. S. (1994). *A evasão do terceiro grau em Curitiba*. Núcleo de Pesquisas sobre Ensino Superior, Universidade de São Paulo.

Rigo, S. J., Cazella, S. C., and Cambruzzi, W. (2012). Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In *Anais do Workshop de Desafios da Computação Aplicada à Educação*, pages 168–177.

Santos, N. V. M. d., Junior, M. L., and Ribeiro, M. L. d. L. (2015). Evasão no curso de engenharia de produção da universidade federal de goiás-regional catalão. In *Anais do ENEGEP*, Fortaleza-CE.