Uso de KDD para Análise do Impacto de Revisões Curriculares

Egídio Loch Terra Karin Becker Cinara Ghedini
Faculdade de Informática – PUCRS – Porto Alegre – Brazil
{egidio, kbecker, cinara}@kriti.inf.pucrs.br

Resumo: *Descoberta de conhecimento em bancos de dados (Knowledge Discovery in Databases - KDD) é o processo não trivial de identificação em banco de dados, de padrões que sejam válidos, novos, potencialmente úteis e compreensíveis. Neste trabalho, abordamos o uso de KDD em um sistema acadêmico, visando determinar e compreender se revisões curriculares trazem alguma forma de prejuízo aos alunos de uma Universidade. Não existe hoje um mecanismo para avaliação do impacto de uma revisão curricular, e KDD apresenta-se como uma alternativa interessante já que o volume de dados envolvido é muito grande, e que, os efeitos nos alunos podem ser bastante individualizados. O presente trabalho descreve um conjunto de resultados obtidos, e analisa a questão sob o ponto de vista de impacto para conclusão do curso. Segundo os resultados parciais, seria prematura a generalização do prejuízo implicado pelas revisões.*

Abstract: *Abstract: Knowledge Discovery in Databases (KDD) is defined as the non-trivial process of discovery of new, valid, potentially useful and comprehensible patterns in large databases. In this work, it is presented an experience on using KDD to identify and understand whether curriculum revisions can affect students in a Brazilian University. Presently, there is no framework to define the notion of impact caused by curriculum revisions, and the use of KDD can bring significant contributions, given the amount of data involved. The present work describes experience results, related to the potential effect of curriculum revision on students requirements for graduation. According to these results, we conclude it is premature to generalize that such revisions affect students negatively.*

1. Introdução

Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases - KDD) é definida em [Fayyad 96] como o processo não trivial de identificação em banco de dados de padrões que sejam válidos, novos, potencialmente úteis e compreensíveis. Esta área tem sido uma das mais estudadas e divulgadas na atualidade, principalmente pelo impacto potencial que a informação extraída pode trazer aos usuários deste processo. A contribuição de KDD em domínios como marketing, vendas, apoio ao cliente, finanças e produção é bastante conhecida [Brachman 96b][Berry 97][Cabena 98][Adriaans 96], e apesar de suas especificidades,

muitos aspectos destas aplicações já se tornaram verdadeiros clássicos, como por exemplo, as regras de associação em "Market Basket Analysis", ou a combinação de agregação e classificação para "customer profiling". No entanto, a aplicação de KDD em novos domínios tem um caráter mais exploratório, na medida em que o levantamento das variáveis relevantes ao domínio e o enquadramento do problema como uma aplicação de KDD são etapas complexas e mais sujeitas a um processo de tentativa e erro.

Este trabalho descreve parte de uma experiência de descoberta de conhecimento em banco de dados em andamento sobre um sistema acadêmico. A base de dados faz parte do Sistema Discente, um Sistema de Informação Legado (SIL) com características temporais que controla todas as atividades de graduação na Universidade Federal do Rio Grande do Sul (UFRGS) - Brasil. O objetivo do processo de KDD é investigar o possível impacto que *revisões curriculares* podem provocar em alunos de graduação da UFRGS.

No Brasil, universidades atualizam seus cursos de ensino superior trocando currículos com conteúdo defasado por currículos mais novos, num processo conhecido como *revisão curricular*. Quando ocorre uma revisão na UFRGS, a universidade tenta evitar que existam currículos simultâneos para um mesmo curso. Isto eqüivale a dizer que deve existir apenas um *currículo vigente* para cada curso em um dado momento. Consequentemente, um aluno pode se formar caso cumpra os requisitos do currículo vigente na época de sua formatura, o qual não é necessariamente o mesmo vigente na época de seu ingresso no curso. O impacto de uma revisão curricular pode ter outras conseqüências na UFRGS, além daquelas relacionadas aos requisitos para a formatura. Considerando que a oferta de vagas não necessariamente atende toda a demanda, esta universidade estabelece uma prioridade para realização da matrícula com base no desempenho acadêmico dos alunos. Esta prioridade é estabelecida em função não somente do desempenho do aluno no semestre anterior (e.g. número de aprovações, reprovações, cancelamentos), mas também em função de seu avanço no currículo (e.g. última etapa completa, de acordo com serialização de disciplinas definida no currículo). Alterações no currículo podem, portanto, prejudicar esta avaliação à revelia do desempenho do aluno.

O problema de compreensão do impacto causado por revisões curriculares foi trazido pelo responsável, há quase uma década, do Sistema Discente, o qual fica sobrecarregado na época de matrícula atendendo alunos insatisfeitos com a maneira como suas prioridades de matrícula são calculadas. A pergunta típica deste alunos é: "como tal aluno pode estar na minha frente, se meu desempenho no semestre passado foi melhor?". A análise de alguns casos em isolado revelou a revisão curricular como principal causa da insatisfação dos alunos, e há interesse em buscar *generalizações*.

A UFRGS oferece hoje mais de 40 cursos, os quais, via de regra, têm sido revisados de maneira freqüente, em média uma vez por ano[1]. Compreender o potencial impacto de uma revisão curricular envolve por

[1] Uma das primeiras descobertas deste processo de KDD.

um lado a análise das características de uma transição de uma versão de currículo a outra, e por outro, as conseqüências destas transições sobre os alunos. Colocado de uma maneira mais prática, uma revisão representa cortes de requisitos antigos e inclusão de requisitos novos dentro do currículo. Alunos podem ser prejudicados, beneficiados ou mesmo não ser afetados pelas alterações destes requisitos de maneiras bastante diversas, e em graus diferentes, em função de seu histórico escolar (e.g. etapa que está cursando, desempenho), ou seja, de modo quase individualizado. Considerando que cada revisão curricular de um curso pode afetar várias centenas de alunos, o uso de KDD se revela como uma das alternativas mais interessantes no processo de generalização dos supostos prejuízos causados pelas revisões curriculares devido ao grande volume de dados envolvido.

O presente artigo tem como objetivo apresentar os resultados obtidos até o momento no processo de caracterização dos efeitos (positivos e/ou negativos) de revisões curriculares na UFRGS com a aplicação de KDD. Como já destacado, experiências de KDD em novos domínios podem ganhar um caráter bastante exploratório [Brachman 96a], e esta tem sido nossa experiência no levantamento e compreensão das variáveis que podem contribuir na caracterização dos efeitos de uma revisão curricular, e no tipo de análise que pode ser realizada sobre elas. Exemplos de aplicações de KDD em sistemas acadêmicos são identificação de padrões de matrícula [Mannila 94], e de afastamento de alunos da universidade [Sanjeev 95] e [Feldman 96].

O restante do texto está organizado da seguinte maneira: na Seção 2 são descritos os aspectos relevantes do Sistema Discente para a compreensão deste trabalho. Na Seção 3 são analisados os critérios para adoção de KDD na base de dados escolhida. Na Seção 4 o processo de KDD em estudo é caracterizado, junto com as alternativas levantadas e os resultados já obtidos. Por fim, a Seção 5 apresenta conclusões e trabalhos futuros.

2. O Sistema Acadêmico

O Sistema Discente é um SIL que controla todas as atividades discentes da UFRGS, sendo composto por três subsistemas: *Alunos*, *Currículos* e *Turmas*. Está em funcionamento há 25 anos, baseado em tecnologia ultrapassada (SGBD Hierárquico e aplicações em COBOL), e não pode sair de funcionamento por servir à missão fim da universidade. Adicionalmente, a base de dados associa um tempo de validade a muitas das informações que representa, introduzindo aspectos de base temporal ao sistema. O subsistema *Currículos* guarda todas as versões de currículos já definidas para cursos oferecidos pela UFRGS, mesmo os já extintos. O subsistema *Alunos* guarda o histórico completo (e.g. admissão, trancamento, matrículas) de todos os alunos e ex-alunos da UFRGS. O subsistema *Turmas* armazena as informações de turmas, como alocação de vagas e salas para estas. Neste trabalho, focamos dois dos subsistemas de Discente, a saber, *Currículos*, por representar revisões curriculares, e *Alunos*, por conter os possíveis prejudicados pelas mesmas.

2.1. Subsistema Currículos

Este subsistema armazena informações sobre todos os curso da universidade, mesmo os já extintos. Alguns cursos possuem ênfases, e os alunos devem escolher a ênfase que desejam seguir ou no momento de seu ingresso, ou em algum ponto antes de sua formatura. O sistema trata todas estas possibilidades de maneira homogênea através de uma identificação única para cada curso ou curso-ênfase, denominada *Codcurso*. No restante deste trabalho, o termo *curso* será usado para representar um curso ou ênfase de curso, sem distinção. A Figura 1 apresenta um diagrama ER estendido dos principais aspectos do Subsistema Currículos, no qual apenas as entidades, relacionamentos e atributos mais relevantes são ilustrados.

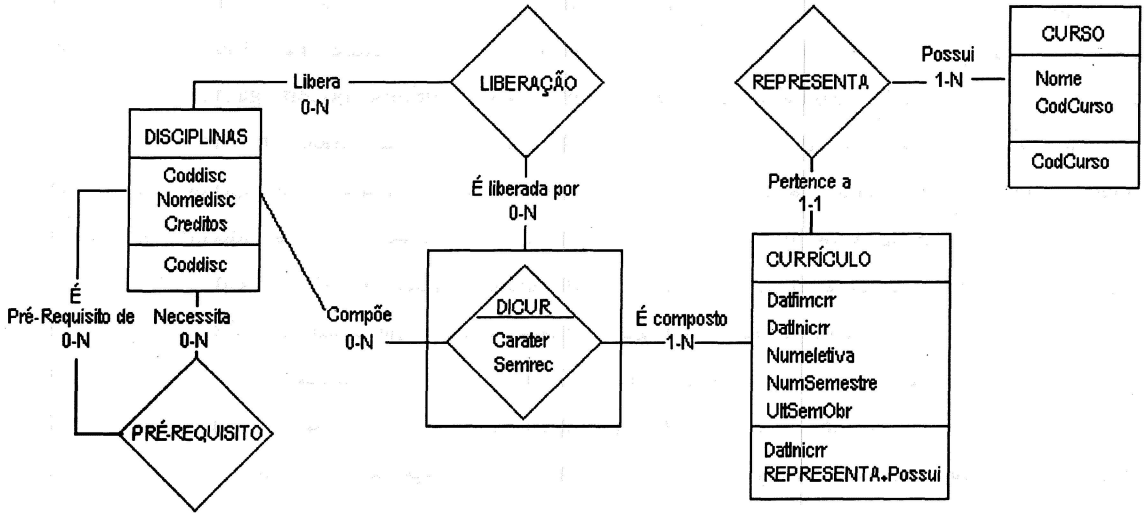

Figura 1. Subsistema Currículos

Cada *CURSO* possui um único currículo vigente em um dado instante de tempo, mas pode possuir diversos currículos se analisado em momentos distintos. Cada *CURRÍCULO* possui um tempo de validade associado (início e fim da vigência- *Datinicrr* e *Datfimcrr*, respectivamente), e é composto por um conjunto de disciplinas distribuídas em semestres. Estas *DISCIPLINAs* possuem propriedades que são independentes do currículo no qual participam, como número de *créditos* e seus *PRÉ-REQUISITOS*, bem como propriedades definidas pelo seu vínculo a um currículo específico (*DICUR*), como seu *Caráter* (e.g. obrigatório, optativa) e o semestre recomendado (*Semrec*). Para facilitar a transição entre currículos foi criado o conceito de *LIBERAÇÃO*, que define regras de aproveitamento de créditos obtidos pelos alunos em currículos anteriores. Uma disciplina de um currículo pode ser liberada por uma ou mais disciplinas que compuseram currículos anteriores do mesmo curso. Com isso, um aluno pode ser dispensado (ou liberado) de uma disciplina se houver cumprido a condição de liberação desta disciplina.

De todas as informações deste subsistema a única que não guarda seu período de validade são os pré-requisitos. Contudo, esta informação não pode ser considerada perdida pois existe na forma escrita na documentação da universidade.

2.2. Subsistema Alunos

O Subsistema Alunos (Figura 2) armazena informação sobre todos os alunos da UFRGS (atuais e ex-alunos). Quando um *ALUNO* ingressa na universidade pela primeira vez, normalmente através de um concurso denominado vestibular, ele recebe um número de *matrícula* único que o identifica na universidade durante toda a sua vida acadêmica. Em situações de transferência de cursos, através de novo concurso, reingresso, etc, seu número de matrícula permanece o mesmo do primeiro ingresso

Para fins de formação, um aluno tem sempre um *VÍNCULO* com algum curso, caracterizado pelo seu início (*Ingresso*) e possivelmente seu final (*Afastamento*). O início e fim são, por sua vez, descritos por uma data e um motivo (*MotIngress* e *Datingress, MotAfasta* e *Datafasta*, para ingresso e afastamento, respectivamente). Em cada vínculo o aluno possui uma *Prioridade* de matrícula composta por vários *índices*, que determinam seu desempenho no curso (número de aprovações/reprovações/cancelamentos, etc) e seu posicionamento entre os alunos deste curso para cálculo de prioridade na matrícula (*OrdemMatr*).

O histórico escolar de um aluno é basicamente composto das disciplinas nas quais ele se matriculou, dos aproveitamentos de créditos obtidos (normalmente quando cursados em outras instituições), e das liberações que obteve em função da equivalência entre disciplinas em currículos distintos. Os dois primeiros casos são registrados em *HISTÓRICO*, junto com o ano/semestre (*Ano*) no qual o aluno cursou/aproveitou a disciplina, o *Conceito* obtido pelo aluno ou alguma observação. No caso das *LIBERAÇÕES*, esta informação não é persistente. Semestralmente, por ocasião da matrícula, uma rotina do sistema determina todas as liberações às quais cada aluno tem direito com base no currículo vigente de seu curso. Para recuperar esta informação, é necessário requisitar ao responsável do sistema a execução da rotina de levantamento de liberações para cada revisão curricular, já que as liberações não são transitivas. Esta rotina tem uma complexidade alta e a título de exemplo, para criar o conjunto de liberações retroativos para um período de 20 anos para apenas dois cursos, foi necessária uma noite inteira de processamento.

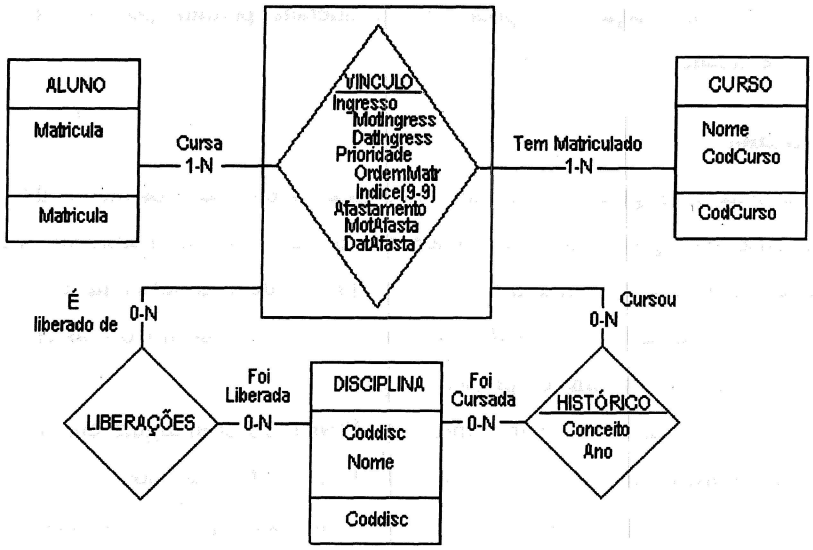

Figura 2. Subsistema Alunos

3. Critérios de Adoção de KDD

Segundo [Fayyad 96], a viabilidade de execução de um processo de KDD em uma base de dados está ligada a um conjunto de critérios, que quando atendidos, aumentam a possibilidade de uma aplicação ser bem sucedida. Estes critérios estão divididos em práticos e técnicos.

Os critérios práticos justificam a aplicação de KDD do ponto de vista administrativo. Entre eles estão:

a) Impacto Potencial da Aplicação: de grande importância para a qualidade do serviço prestado pela universidade, pois o impacto de novas revisões pode ser então medido até mesmo antes de sua implantação, possibilitando eventuais ajustes;

b) Falta de Alternativa: não existe hoje um mecanismo para avaliação do impacto de uma revisão curricular. Além disso, o volume de dados envolvido é muito grande, o que inviabiliza uma análise manual de um conjunto expressivo de dados, já que os efeitos nos alunos podem ser bastante individualizados. Logo a aplicação de KDD coloca-se como uma das alternativas mais interessantes para o problema;

c) Suporte Organizacional: primordial desde o início da aplicação. Foi o responsável pelo sistema que levantou a hipótese de realização deste trabalho pelo interesse não só da universidade mas como também da equipe responsável pelo sistema, e tem colaborado desde o início para o bom andamento do projeto;

d) Problemas Legais: todas as informações que permitem identificar alunos foram criptografadas.

Existem também os critérios técnicos, que justificam, do ponto de vista da informação presente, a viabilidade do processo. Entre os critérios estão :

a) Quantidade de Dados e Atributos Relevantes: a quantidade de informações existente no Sistema Discente é alta, e vem se acumulando desde seu início, na primeira metade da década de 70. Hoje, o número de cursos armazenado neste sistema, bem como a quantidade de currículos que já foram utilizados é muito grande. Além disso, estão registrados no sistema mais de 100.000 alunos, cujas informações de histórico chegam ao total de 4.000.000. Esta grande quantidade de dados e complexidade de relacionamentos privilegia o uso de KDD em detrimento a técnicas de análise manuais. Existe uma grande quantidade de atributos neste sistema, no entanto, existem poucos atributos que caracterizem diretamente o impacto de uma revisão curricular, já que a definição de impacto, e levantamento de seus fatores é um dos objetivos desta aplicação. O que se tem feito é derivar métricas para avaliação de impacto a partir de atributos existentes, como a contagem de requisitos necessários à formatura antes e depois de uma revisão curricular. Este tópico é discutido na Seção 4 em mais detalhe;
b) Qualidade dos Dados: não se tinha informações precisas sobre a qualidade dos dados no início deste trabalho, mas segundo [Adriaans 96], a quantidade de ruído não pode ser avaliada sem que as informações na base de dados sejam utilizadas. Logo, o critério de qualidade não pode inviabilizar uma análise ainda em sua fase inicial, pois sua avaliação é restrita no início da aplicação.
c) Conhecimento acerca do domínio : necessário para criação de intervalos de confiança e possibilidade de execução da análise. Com a interação com o especialista e pessoas envolvidas com o sistema é possível validar os resultados extraídos com maior segurança. Se os resultados forem estranhos então é possível interagir para verificar sua validade.

Baseado na análise dos critérios práticos e técnicos, considera-se o processo de KDD para averiguação do impacto das revisões curriculares viável e adequado.

4. O Processo KDD

Um processo KDD é altamente interativo e iterativo [Fayyad 96][Brachman 96a][Adriaans 97], sendo subdividido em várias etapas. De uma maneira mais geral, estas etapas podem ser agrupadas em três grandes fases: preparação, análise e interpretação. Grande parte das pesquisas nesta área estão centradas na tarefa de análise, mais especificamente na mineração de dados, que trata da aplicação de métodos sofisticados de análise estatística e de aprendizagem automática a fim de buscar padrões sobre um grande volume de dados. Contudo, estima-se que a mineração de dados propriamente dita consome apenas 15 a 20% de todos os esforços do processo [Brachman 96b]. Na prática, boa parte do processo é centrada na fase de preparação, que inclui a

compreensão, seleção, transformação e limpeza dos dados, a fim de enquadrar o problema existente no domínio como um problema de mineração de dados, e definir as variáveis e dados relevantes para o processo de análise.

Na fase de preparação, a caracterização do Sistema Discente como um SIL acarretou problemas sérios de qualidade. Assim, a fase de preparação foi constantemente repassada quando era detectado um novo problema de qualidade [Terra 99], caracterizando a iteratividade e interatividade do processo. Os problemas de qualidade e a novidade do domínio nos fizeram optar por uma abordagem incremental para a caracterização da análise do impacto trazido por revisões curriculares. O presente trabalho representa os primeiros resultados obtidos, tanto em termos de definição de fatores/variáveis que caracterizam o impacto, quanto no que tange os dados analisados, e representam apenas o primeiro esforço na tentativa de encontrar padrões válidos nos efeitos trazidos por revisões curriculares. O processo completo envolvendo estes resultados, bem como os problemas de qualidade causados pelo SIL, são discutidos em detalhe em [Terra 99].

4.1. Variáveis para Medir o Impacto

Como já mencionado, não existe hoje uma definição de como medir o impacto de uma revisão curricular. Em princípio, os alunos são afetados se a revisão implica dificuldades para a conclusão do curso (e.g. aumento do tempo de permanência), ou dificuldades para realização da matrícula (e.g. afeta atributos usados para cálculo das prioridades), sendo que os dois problemas estão interligados. No Sistema Discente existem muitos poucos atributos que caracterizem diretamente estes dois aspectos, sendo necessário criar formas de medí-los. Na presente fase da pesquisa, principalmente pela dificuldade de conseguir dados consistentes, foi possível abordar apenas o primeiro aspecto, isto é, conseqüências para conclusão do curso.

Para derivar o valor de impacto criamos uma medida relacionada com os alunos (que sofrem prejuízo ou benefício quando acontece uma revisão) e com a forma da revisão (o que foi alterado na transição curricular). Se uma revisão curricular altera os requisitos necessários para obtenção do grau do curso que este currículo representa, é necessário saber quais são os requisitos que foram alterados, inseridos e removidos de um currículo para o seu subsequente. Estas modificações podem ou não afetar os alunos que estão matriculados no curso e que no momento da revisão ainda não se formaram, e é neste que ponto entram as informações do histórico dos alunos. Examinando cada aluno individualmente, é possível saber se as modificações curriculares vão ou não afetá-lo, mas é necessário que sejam analisados diferentes alunos, pois uma revisão curricular pode afetar um grupo de alunos e não afetar outro, ou mesmo afetar de formas diferentes.

Para criar o conjunto de dados foram contados os números de disciplinas aprovadas (*ndisc1* e *ndisc2*) e de liberações (*nlib1* e *nlib2*) que cada aluno tinha antes e depois de cada revisão em relação a cada currículo

vigente[2]. À diferença das contagens obtidas antes e depois da revisão chamamos ***impacto***. O conjunto de dados criado para análise possui as seguintes informações (Tabela 1) : aluno (*Matricula* - criptografada), ano/semestre da revisão (*Revisão*), contagens (*Ndisc1, Ndisc2, Nlib1* e *Nlib2*), medida de impacto (*Impacto*) e o ano de ingresso e de formatura do aluno (*Ingresso* e *Formatura*). A partir deste conjunto de dados, o valor para novas variáveis puderam ser extraídas como, por exemplo, o tempo de permanência do aluno no curso (*Tperm* entre *Ingresso* e *Formatura*), o número de revisões curriculares pelas quais cada aluno passou (*Ncur)* ou então qual o impacto total do aluno em toda sua história no curso (*Soma_impacto)*, que é a soma de impacto em cada uma das *ncur* revisões curriculares sofridas pelo aluno). Várias outras variações de impacto também foram definidas, tais como impacto considerando somente disciplinas obrigatórias, percentagem de conclusão do curso, entre outras.

Matrícula	Ingresso	Formatura	Revisão	Ndisc1	Ndisc2	Nlib1	Nlib2	Impacto
10392183	1978-0	1984-0	1980-0	23	21	3	2	3
10392183	1978-0	1984-0	1980-5	23	24	2	3	-2
38210293	1983-0	1989-5	1988-0	14	13	10	11	0

Tabela 1. Tabela Impacto com alguns exemplos

A complexidade existente nas relações entre as informações armazenadas no Sistema Discente permite a criação de vários outros conjuntos de dados que poderiam ser relacionados com o prejuízo causado por uma revisão curricular. Por isso, consideramos que as variáveis utilizadas nas análises discutidas no presente trabalho constituem apenas um conjunto inicial. No entanto, as dificuldades encontradas para extrair estes dados fazem com que o processo se prolongue sem que haja uma convicção a respeito de sua real representatividade na caracterização do impacto causado por revisões curriculares, o que reforça a condição exploratória desta aplicação. Mesmo com estas considerações, achamos que o conjunto criado para estudar o impacto das revisões é significativo.

4.1.1 Métodos de Análise

Em [Agrawal 93] é feita uma primeira tentativa de dividir métodos de mineração em grupos chamados de classes de problemas. Em [Fayyad 96] outra divisão é apresentada, onde os métodos são agrupados em tarefas de mineração. Não existe uma divisão considerada unânime para todos os métodos existentes, tampouco ortogonal, isto é alguns métodos fazem parte de um grupo ou de outro, dependendo do autor. Por esta razão não entraremos em maiores detalhes e consideraremos os seguintes métodos : sumarizações, associações,

[2] Para contar as liberações é necessário executar uma rotina do sistema que calcula *LIBERAÇÕES* (Figura 2) para todo semestre anterior a cada revisão analisada, já que essa informação não é persistente, o que consome um enorme tempo de processamento.

classificações, agrupamentos e regressões. Para cada um método existem vários algoritmos que podem ser utilizados.

Escolher um método apropriado para resolver o problema em análise nem sempre é uma tarefa fácil. Para aplicar um algoritmo referente a qualquer um dos métodos, é necessário ter algum tipo de conhecimento a respeito de sua interface, ou seja, como ele requer que sejam formatados os dados, como são apresentados os resultados (como deve ser sua interpretação) e, principalmente, quais parâmetros (utilizados para guiar o processo de extração) devem ser ajustados e de que maneira. Além disso, os métodos determinam como serão tratados os problemas em análise, o que nem sempre é muito claro na aplicação. Nesta experiência, por falta de aplicações prévias que auxiliassem esta tarefa, não foi possível saber se o melhor método a ser utilizado seria agrupamento, classificação, associações ou sumarizações. O único método descartado inicialmente foi o de regressões, já que não estávamos querendo criar um modelo para prever um comportamento futuro e sim criar um modelo para compreendê-lo, e não repeti-lo mais em casos de prejuízos.

A sumarização pode ser utilizada para obtermos descrições compactas do conjunto de dados. Na nossa experiência, podemos utilizar este método para verificar a correlação entre impacto e tempo de permanência, ou entre tempo de permanência e número de revisões curriculares. Além disso, também é possível fazer cálculos mais simples, como por exemplo somar o impacto de todos os alunos em uma revisão curricular para identificar quais revisões foram mais ou menos prejudiciais.

As associações são utilizadas para identificar quais valores de atributos distintos ocorrem freqüentemente em conjunto. Por exemplo, com regras de associações podemos tentar identificar quais revisões, impactos acontecem em ingressos comuns (i.e. alunos de uma mesma turma). Assim, poderíamos detectar a condição de prejuízo/benefício de um impacto em relação à condição dos alunos que passaram pela revisão.

Os agrupamentos podem ser utilizados para criar conjuntos de elementos que possuam similaridades através dos atributos escolhidos. Uma forma de aplicar um agrupamento em nossa experiência seria com uso dos atributos impacto, revisão, ingresso, tempo de permanência e ano de conclusão para dividi-los em grupos com similaridades.

As classificações podem ser utilizadas para caracterizar classes previamente descritas. Uma forma de utilizar as classificações em nossa experiência seria identificar quais revisões foram mais prejudiciais, através da criação de agrupamentos ou sumarizações, e, em seguida, caracterizá-las por exemplo pela quantidade de disciplinas e liberações que os alunos tinham antes e depois dela.

4.1.2 Resultados

Para realizar as análises foram utilizadas várias ferramentas que permitiram desde a extração dos dados até a visualização de resultados obtidos. Estas ferramentas compreendem : SGBD Oracle 8, linguagem de

consulta (SQL, PLSQL), planilha eletrônica com funções estatísticas (e.g. correlação, médias) e 3 ferramentas de KDD: BKD [Ramoni 97], DBMiner [DBM 97] e Clementine[Integral 98].

A execução da análise mostrou resultados inesperados nos cursos que foram utilizados como conjunto de dados, a saber, curso 101-00 e 102-00. Para a análise aqui descrita foram selecionados apenas alunos formados, para que fosse possível utilizar o tempo de conclusão do curso (*tperm*) como uma das variáveis de análise. A interpretação dos resultados obtidos nos leva a considerar ***prematura*** a hipótese que generaliza o prejuízo causado por revisões curriculares. É necessário que sejam examinados mais cursos e mais variáveis. Contudo o difícil processo de extração e validação de dados dificulta a rápida construção de um novo conjunto de dados confiável[3]. Os resultados obtidos até agora são parciais, mas demonstram a dificuldade em tratar todos cursos de forma homogênea, já que cada curso possui características próprias de organização de currículos e tem reflexo desta organização em alunos com comportamentos não necessariamente iguais.

A dificuldade inicial de estabelecer quais atributos interferem ou representam impacto levou à criação de hipóteses a respeito de impacto. Como descrito na Seção 4.2, foram criadas fórmulas para calcular o valor de impacto para todos alunos em todas as revisões pelas quais estes alunos passaram. Entre as várias formas de análise aplicadas houveram várias descobertas. Muitas delas foram encontradas durante a fase de preparação dos dados e confirmação de conceitos que as pessoas envolvidas tinham no Sistema Discente. A qualidade ruim e o longo tempo de operação da base (com evolução de conceitos) geram uma desconfiança acerca dos dados armazenados. Esta desconfiança pode ser verificada com um processo de validação dos conceitos através dos dados existentes [Terra 99].

No levantamento do número de currículos dos cursos da universidade, foi verificado que a quantidade de revisões é altíssima, e com pouca variação entre todos os cursos, que têm médias próximas a 1 revisão por ano. Este resultado contradiz o conceito inicialmente passado pelos responsáveis pelo sistema de que existiriam alguns cursos mais estáveis e outros menos (i.e. com revisões mais ou menos freqüentes). Este resultado põe em dúvida a possibilidade de diferenciar os impactos de revisões curriculares nos alunos pela *quantidade* de revisões, e aumenta a possibilidade de uma procura *qualitativa* por revisões mais prejudiciais. Em outros termos, existem algumas revisões curriculares que são piores (ou melhores) que outras.

A hipótese levantada pelo responsável do sistema é que seria possível generalizar que alunos são prejudicados pelas revisões curriculares. Uma forma de validar esta hipótese é verificar se existe algum relacionamento entre tempo de permanência (*Tperm*) e número de revisões curriculares (*Ncur*). Se houver um relacionamento é possível aprofundar a análise para descobrir como a revisão afeta o tempo de permanência.

[3] A título de exemplo, a construção da tabela Impacto para os dois cursos citados levou vários meses, até que fossem obtidos dados limpos.

Este relacionamento foi avaliado segundo uma análise de dependência estatística entre as variáveis: a correlação. O resultado encontrado foi uma alta correlação entre os tempos de permanência dos alunos e seus números de revisões curriculares. Isto significa que quanto mais um aluno fica na universidade por mais currículos ele passará, mas também pode ser interpretado ao inverso: quanto mais revisões atingirem um aluno maior será o seu tempo de permanência no curso.

Pela correlação não é possível determinar qual variável influencia a outra, e para detalhar esta análise foi utilizado um modelo de redes bayesianas [Ramoni 97], que permite identificar se existe ou não dependência, e qual o sentido desta. O modelo criado pela ferramenta BKD mostrou que não é possível na relação de dependência determinar qual variável determina o valor da outra, mas confirmou a existência de dependência entre elas.

A inclusão da variável *soma_impacto*, que totaliza para cada aluno o impacto que este sofreu em todas revisões pelas quais passou, permite enriquecer esta análise pois é possível relacioná-la a *tperm* para verificar se o impacto gerado no aluno fez com que este demorasse mais para concluir o curso. Os resultados foram levantados através de medida de correlação, e os valores conseguidos através dessa análise nos cursos utilizados apresentaram valores de correlação baixíssimos : 101-00 (correlação –0,001) e 102-00 (correlação –0,038). Estes resultados praticamente desbancam a hipótese de que, nestes cursos, revisões curriculares aumentam os tempos de permanência dos alunos. Em alguns casos, alunos que tiveram impacto alto obtiveram tempo de permanência baixo, o que contradiz a hipótese levantada.

Com a falta de comprovação da hipótese quantitativa, o estudo de revisões específicas permite que sejam compreendidas as razões de maior ou menor impacto gerado por estas revisões. Contudo, a distribuição dos valores da variável *impacto* também mostram pouca potencialidade em termos de conhecimento que possa ser extraído. Isto porque a grande maioria dos alunos analisados teve *impacto 0*, ou seja, não foi afetada segundo as métricas levantadas. Este é um dos motivos da baixa média de *impacto* por revisão em todos os alunos dos dois cursos (101-00 teve 1,91 e 102-00 teve 0,09). Ainda assim, em algumas revisões o impacto é destacado das demais pela alta quantidade de alunos atingidos ou pela quantidade de perdas destes alunos. Assim, pudemos identificar quais revisões podem ser analisadas em maior detalhe para verificar se suas características podem ser generalizadas como causadoras de prejuízo.

Para generalização do modelo apresentado pelas revisões mais prejudicais foram consideradas técnicas de agrupamento e classificação disponíveis na ferramenta Clementine. Para agrupamento existem dois métodos : K-Means e Redes Neurais (Kohonen). No método de K-Means, a distância de um ou mais pontos centrais para todos elementos serve como medida para criar os agrupamentos. A utilização de uma média faz com que este método seja mais indicado para problemas onde as variáveis são numéricas e não categóricas, como era nosso caso. O outro método de agrupamento é os mapas de Kohonen. Estas redes neurais fazem um mapeamento entre

duas dimensões (entrada e saída), objetivando encontrar uma representação compacta. Após a criação de um mapa de Kohonen é necessário fazer sua interpretação visual, a fim de selecionar os conjuntos que aparecem agrupados e gerar um conjunto a parte onde estarão somente os exemplos que o representam. Com este novo conjunto podem ser geradas novas representações que permitem compreendê-los. Os dados podem ser categóricos ou numéricos.

Os resultados obtidos com os dois métodos foram semelhantes e têm relação com a quantidade de informação que pode ser extraída no conjunto de dados selecionados para análise (101-00 e 102-00). A existência de uma grande quantidade de alunos com impacto 0 (zero) faz com que todos os agrupamentos gerados incluam este impacto como predominante, o que não é um resultado interessante se considerarmos que a hipótese de haver prejuízo deve ser melhor investigada.

O mesmo problema apresentado nos agrupamentos se repete nas classificações efetuadas, ou seja, são poucos os caso onde o impacto é diferente de 0 (zero). Ainda sim, foram efetuadas análises de classificação utilizando três atributos : o *impacto* de cada aluno em cada revisão, o *ingresso* de cada aluno e o ano de *revisão*. Nesta classificação, é necessária a escolha de um atributo que será tratado como o definidor da classe. Foram utilizados os três atributos, alternadamente, para perfazer o papel de classe, no entanto o único que apresentou resultados que pudessem ser úteis foi o atributo *revisão*.

Com base na sumarização realizada, o ano de 1988 teve a *revisão* cujo maior número de alunos do curso 101-00 foi atingido de forma negativa (prejuízo). Com esta definição foram encontradas regras onde aparecem os *ingressos* de 1983, 1984 e 1985 como sendo os que sofreram um prejuízo padronizado em todos os alunos (*impacto* 1 ou 2). Apesar deste resultado, pouco foi acrescentado ao que já se havia descoberto - a pior revisão é de 1988 - e grande parte dos alunos que passaram por ela ingressaram a partir de 1982.

Outras tentativas também foram realizadas, como por exemplo, a criação de regras de associação para as variáveis escolhidas usando DBMiner, mas em todas as regras encontradas a precisão (suporte) era muito baixa, devido aos problemas de distribuição de impacto (i.e. predominância de impacto 0).

A conclusão é que os resultados obtidos são específicos demais para serem generalizados, ou seja, as revisões curriculares não podem ser consideradas prejudiciais sem que sejam considerados os estados dos alunos no momento em que elas ocorrem. Assim, mesmo que existam revisões semelhantes entre si, seus impactos nos alunos (prejuízos ou benefícios) dependem do estado destes no momento em que elas ocorrem. É cedo para tentar generalizar os resultados encontrados até agora. Certamente outros cursos deve ser analisados usando os procedimentos aqui descritos. Além disso, outros fatores ainda não considerados também podem trazer algum esclarecimento adicional em termos de resultados. No entanto, mesmo com um procedimento já definido para criação de novos conjuntos de dados utilizando outros cursos é difícil fazer a aquisição dos dados pois muitas

informações são calculadas através de rotinas (e.g. liberações em *LIBERAÇÕES*), a qualidade é baixa, e alguns dados que poderiam ser úteis não estão disponíveis de forma digital (e.g. pré-requisitos em *PRÉ-REQUISITOS*).

5. Conclusões

A área de KDD é recente e ainda possui vários desafios a serem superados. Novos domínios ainda não explorados tornam as aplicações muito distintas daquelas onde já existem pontos consagrados. Isto porque o processo de KDD possui muitos detalhes que devem ser cuidadosamente acertados para que resultados positivos sejam extraídos, e a experiência com aplicações clássicas já permitiu a elaboração de um *know how* em como fazer ajustes. Em aplicações exploratórias, é necessário um alto conhecimento do domínio em questão e um bom conhecimento dos métodos de análise existentes. Estes dois requisitos são necessários para que sejam alcançados resultados satisfatórios, mas não são suficientes. Também é necessário ter um conjunto de dados adequado à solução do problema, sob pena de uma boa aplicação do processo resultar em uma série de resultados inexpressivos.

Neste trabalho aplicamos este processo em uma base acadêmica – o Sistema Discente – com o objetivo de estudar um problema proposto pelo seu responsável. A viabilidade de aplicação foi verificada segundo alguns critérios que justificam o uso de KDD, e o processo iniciado com uma hipótese que deveria ser confirmada e compreendida: o impacto sofrido por alunos causado por revisões curriculares.

A conclusão até o presente momento é que é prematura esta tentativa de generalização do prejuízo. A complexidade do domínio e a falta de qualidade fizeram com que a aplicação se estendesse por muito tempo, dificultado de levantamento de novas variáveis, dados confiáveis e realização de novas análises que permitissem chegar a resultados mais conclusivos. Em nossa aplicação, vários métodos de análise foram utilizados para entender o problema proposto, ocasionando várias iterações dentro do processo. Estas iterações foram paralisadas no momento em que detectamos que o conjunto de dados não poderia fornecer conhecimento além do que já havíamos conseguido. Contudo, o conjunto utilizado pode ser estendido com a incorporação de novas informações ou novos conjuntos podem ser criados para continuação deste processo que pode ser continuado *ad infinitum*. A aplicação de KDD em sistemas legados e temporais é nova, e os problemas encontrados em nossa aplicação devem ser generalizados para outras aplicações semelhantes.

Como trabalhos futuros, além da aplicação das mesmas análises em outros cursos, pretende-se abordar fatores que caracterizem o efeito de revisões curriculares na prioridade de matrícula dos alunos.

6. Bibliografia

[Adriaans 96] Adriaans, P. & Zantige, D. **Data Mining**. Harlow. Addison-Wesley, 1996.

[Agrawal 93] Agrawal, R.; Imielinski, T., Swami, A. Database Mining : A performance perspective. In: **IEEE transactions on Knowledge and Data Engineering**, New York, IEEE Computer Society. vol. 5 n° 6 Dezembro 1993.

[Berry 97] Berry, M. & Linoiff, G. **Data mining techniques for marketing, sales and customer support**. New York (NY). John Wiley & Sons, 1997.

[Brachman 96a] Brachman, R. J., Anand, T. *The Process of Knowledge Discovery in Databases: A Human-Centered Approach,* In : **Advances in Knowledge Discovery and Data Mining**. Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (Eds.). Menlo Park (CA). AAAI, 1996. p. 37-57.

[Brachman 96b] Brachman, R. J. *et alli. Mining Business Databases.* **Communications of ACM**, New York (NY), ACM. Vol. 39, no. 11, Novembro 1996. p. 42-48

[Cabena 98] Cabena, P. et al. **Discovering data mining: from concept to implementation**. New Jersey (NJ). Prentice Hall, 1998.

[DBM 97] **DBMiner : A System form Mining Knowledge from Large Data Sources**. Data Mining Research Group, School of Computing Science, Simon Fraser University, British Columbia, Canada. Capturado em 22 Jul. 1998. Online. Disponível na Internet em http://db.cs.sfu.ca/DBMiner.

[Feldens 96] Feldens, M. A.; Castilho, J. M. V.. Descoberta de Conhecimento Aplicada à Detecção de Anomalias em Base de Dados. Em : Conferencia LatinoAmericana de Estudos em Informatica. Santafé de Bogotá, 1996.

[Fayyad 96] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery : An Overview. In : **Advances in Knowledge Discovery and Data Mining**. Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (Eds.). Menlo Park (CA). AAAI, 1996. p. 1-34.

[Integral 98] Integral Solutions Limited. **Clementine User Guide Version 5**. Integral Solutions Limited. [sl],1998.

[Mannila 94] Mannila, H.; Toivonen, H. ; Verkamo, I. **Improved Methods for Finding Association Rules**. Internal Paper. Universidade de Helsinki. Finlandia 1994

[Ramoni 97] Ramoni, M; Sebastiani, P. **Learning Bayesian Networks on Incomplete Databases**. Kmi-Techincal Report Kmi-TR-41, 1997. Capturado em 22 Jul. 1998. Online. Disponível na Internet em http://Kmi.open.ac.uk.

[Sanjeev 95] Sanjeev, A.P.; Zytkow, J.M. Discovering Enrollment Knowledge in University Databases. In : Proceedings of the 1st International Conference on Knowledge Discovery & Data Mining. No.1, 1995 Montreal. Proceegings... Menlo Park (CA); AAAI, 1995. p. 242-251.

[Terra 99] Terra, E. **Uma Experiência de Descoberta de Conhecimento em uma Base de Dados Legada e Temporal**. Porto Alegre, 1999. 103p. Dissertação (Mestrado em Informática) – Faculdade de Informática, PUC-RS, 1999.